CORRELATION THEORY

In the vernacular, if two variables are correlated, then they are somehow related. In scientific discussion the term correlation has a more limited and specific meaning. If two variables are said to be correlated, this means that both variables can be characterized by real or complex numbers. It also implies that either of the variables can be used to predict the other. More specifically, the prediction can be accomplished by a linear function. For example, the variable θ and the function $\cos\theta$ are uncorrelated. The function can be predicted exactly from the variable, but the prediction method is not linear.

These (usually unspoken) assumptions are implicit in all correlation analyses: numerical representation, linear prediction, and judgment of the prediction by a least-squares, or root mean square (rms), criterion.

The problem of finding a numerical representation for the data is not always easy. Most of the following discussion will assume that the data originate as time-dependent waveforms. However, these long strings of numbers are rarely immediately useful. There are too many of them, and most of them contain little or no useful information. The first task is usually to extract parameters from these waveforms. Often correlation analysis is used first to extract the parameters and then to analyze the parameters.

It is always legitimate to question whether only linear functions should be considered. In many cases there is good reason to believe that a nonlinear function is appropriate. The problem of fitting a nonlinear function to a data set is not always more difficult than fitting a linear function. The difficulty is that there are no standard techniques that lend themselves to routine use. Often it is necessary to devise a new approach for each problem. For example, suppose that one needs to fit a nonlinear function of x to a data set y over a specific interval by adjusting parameters a, b, and c. It may happen that a dominates the function over a portion of the interval and then becomes unimportant over the rest of the interval. Similarly, b may have little or no effect on the peak value of the function but control the rate of decay of the function. One must first examine the function to see what effect each parameter has, and then perhaps adjust them separately. This may be fairly easy to do by inspection, but difficult to automate. By contrast, linear functions give equal importance to each variable and to all parts of the interval (although it is possible to weight some parts of the interval more heavily than others in many cases).

The insistence on linear functions for prediction is sometimes compromised. For example, a polynomial may be treated as a linear sum of powers of the variable. Other nonlinear functions may be inserted into the summation with linear multipliers.

The choice of an rms criterion for judging the quality of a predictor is not necessarily obvious and should not always be taken for granted. In some cases it may be more important to minimize the worst-case error than the rms error. This criterion leads to a minimax problem. In a few cases the average of the absolute value of the error may be a better criterion. This may lead to median estimators.

However, the rms criterion has proven to be by far the most fruitful assumption in the majority of cases. This is due largely to the ease with which second moments of rms solutions can be followed through linear transformations of the variables.

It is important to understand that correlation analysis can never prove a cause-and-effect relationship. If two variables are causally related, correlation analysis cannot determine which variable is the cause. Often both variables are caused by a third variable that is not even known. About the only thing that can be said with certainty is that if two variables are independent then they are uncorrelated.

However, correlation analysis is an important tool to study relationships of many types. That two variables are uncorrelated is not reason enough to dismiss the possibility that they are related. But if they are correlated it is reasonable to try to figure out why. Also, the absence of a predicted correlation can lead to important discoveries.

PHYSICAL MEASUREMENTS AND THE DECIBEL SCALE

In this discussion, the key quantities of interest are often ratios of averages of squares of variables. In physical systems, the variables often are voltages or pressures or flux densities. In these cases the physical power is proportional to these mean squared values. Discussion of these power levels often involves several problems.

First, the quantities often vary over ranges that are difficult to imagine. The quietest sound that a human can hear corresponds to a pressure on the order of 0.00002 Pa (rms), while the sound level on a jet plane can be over 5 Pa. Second, the uncertainty of the measurement often varies with the level. A good acoustic measurement may have an uncertainty of 20%, making measurements at low levels appear much better than measurements at higher levels.

The usual solution to these problems is the decibel (dB) scale. The decibel scale is a logarithmic scale. However, a common logarithmic scale is felt to be a bit too coarse, so the logarithm is multiplied by 10. The key to understanding the decibel scale is to remember that by convention it is always a ratio of power, or energy values. Suppose the rms acoustic pressure in a room is 0.2 Pa. An acoustician, remembering that acoustic power goes as the square of the rms pressure, might compute $10 \log(0.2^2/0.00002^2) = 80$ and say, "The sound level in the room is 80 dB relative to 0.00002 Pa." Of course, he or she could get the same answer by computing $20 \log(2/0.0002)$, so it is often said that the sound level goes as 20 times the logarithm of the pressure.

In some instrumentation problems it can be tricky to keep track of whether the values should be plotted as 10 log or 20 log. The key to keeping it straight is to ask, "How would the quantity behave if the power were doubled?"

In the same vein, the decibel scale says nothing about the units of measurement. The reference to 0.00002 Pa is a specification of a physical state, not a system of units. If one cannot relate the variables to a power level, then the decibel scale is not appropriate. Engineers will occasionally make comments like "His salary went up by 1 dB when he got the promotion." The implied humor is that the speaker is also saying "Money is power."

TRANSFORMATIONS

Modern data collection problems tend to involve great amounts of data, most of which have no value. It is important to select the small subset of the data that is of potential value. The solution is to attempt to transform the data in such a way as to concentrate the important information into a small number of parameters. The most effective way to do this is usually with the Fourier transform.

In its most common form, the Fourier transform represents the data as a summation of sine and cosine waves, or complex exponentials. Other function sets are sometimes used, (e.g., Walsh functions or Bessel functions), but not often. The reason for the preeminence of complex exponentials as basis functions is the ease with which time translations are handled. Often the data look the same from one time to another and absolute time has no physical significance in interpreting the waveform. (This assumption is referred to as *time stationarity*. Various types of stationarity are defined, depending on how rigorous a concept of time stationarity is needed, but the general idea is that it is impossible to infer absolute time from the waveform.) Even if the waveform is impulsive, it is confusing if its representation changes drastically with arbitrary shifts of the time origin, as happens with some of the alternatives to the complex exponentials.

This invariance with respect to the start time of the signal gives rise to another important aspect of the complex exponentials. They look the same after they have been operated on by a linear filter (i.e., the complex exponentials are eigenfunctions of linear differential equations). This is the key idea. If a summation of complex exponentials is passed through the linear filter, the filter may amplify or delay each component by a different amount, but it does not mix the different frequencies. The output at each frequency depends only on the input at that frequency and is independent of any other frequency.

Thus, the most useful approach known for studying how a waveform will change as it passes through a linear filter is as follows: First the waveform is represented as a summation of complex exponentials at various frequencies. Then the effect on each frequency is calculated separately to see how its amplitude and phase will change as it passes through the filter. Finally, the altered exponentials are summed to give the waveform that emerges from the filter. Since so much of our world is governed by linear differential equations, the importance of understanding waveforms in terms of their Fourier representation is difficult to exaggerate.

The Fourier transform is often best understood by thinking of it as a frequency shift followed by a low-pass filter. For a frequency of interest, the waveform is frequency-shifted by multiplying it by the sine and cosine waves, or the complex exponential wave. This shifts the information of interest to the band around zero frequency. The waveform is then low-pass-filtered to eliminate information at other frequencies, and the result is the Fourier coefficient that describes the waveform at the frequency of interest.

It is tempting to believe that the complex exponentials are only mathematical artifacts, while reality is restricted to the wave as it is represented in time. Common experience refutes this. For example, an AM radio receives an electrical signal that is mostly meaningless noise. The circuitry then separates out the real information for the station of interest and sends the resulting Fourier coefficients to the speaker to produce meaningful sounds (plus advertisements and political commentary).

The Fourier coefficients are usually represented by complex numbers, so the theory of complex numbers is intimately connected with many engineering problems.

The terminology of complex variables is misleading and unfortunate. The name "complex numbers" suggests that they are difficult to handle. In fact, complex variables are popular because their use greatly simplifies many problems. The terminology also suggests that the "imaginary" part of the complex variable is somehow less intimately connected with reality than the "real" part. This idea is dangerously wrong. For example, electrical circuits sometimes develop large imaginary voltages. These imaginary voltages can cause arcing between supposedly isolated parts of the circuit. They can break down circuit components, and they can kill unwary handlers.

In the following discussion, the term *analytic* will be used to describe a function that is analytic in the sense of complex analysis theory. Any of several equivalent definitions may be used. For example, a function is analytic if it has a derivative in the ordinary sense, or if it is the derivative of another function, or if it has a power series (Taylor series) representation. In the same vein, an *analyticity* is a point at which a function is analytic. When a function is analytic that fact has profound implications, most of which are far beyond the scope of this article.

THE DETECTION PROBLEM

Correlators are often used to decide the presence or absence of a particular signal. The investigator begins with a wave-

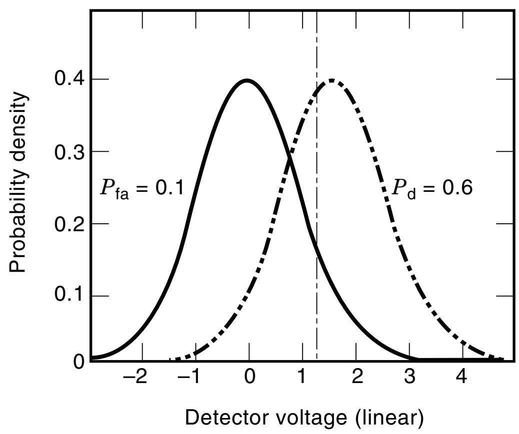

Figure 1. Probability density curves for noise only (left) and signal plus noise (right) for matched filter detector. The threshold is chosen for a false-alarm rate of 10%. The detector characteristics are determined entirely by the ratio of the horizontal separation (the signal strength) and the standard deviation.

form that may or may not contain the signal. By correlating the waveform with the signal, a value is obtained whose statistics depend on the amount of signal energy in the waveform. The analytic details will be discussed below, but the reasoning used for the test is illustrated in Fig. 1.

Figure 1 shows the probability density function for the correlator output when the signal is absent and when it is present. In each case, the probability density function is bell-shaped. Sometimes the function is truly Gaussian, and sometimes it can be approximated as Gaussian. (This approximation is sometimes dangerous, as will be discussed below.) A threshold has been established, which is indicated by the vertical line, and if the correlator output is above this threshold the equipment is to issue an alarm signal.

The probability of a false alarm is the area under the left curve that is to the right of the threshold. In the illustration, a threshold has been set to provide a 10% probability of a false alarm (P_{fa} = 10%). That is, when no signal is present, the correlator will produce an alarm 10% of the time. The signal strength is represented by the horizontal separation of the two curves. When the signal is present, the probability of detecting it is the area under the right curve that is to the right of the threshold. In this case, the signal energy is strong enough that, when present, it will be correctly detected 60% of the time. This is the probability of detection (P_d = 60%). The areas under the curves to the left of the threshold give the probability of a correct dismissal (P_{cd} = 90%) and the probability of a miss (P_m = 40%).

Figure 1 also illustrates several other important concepts. The noise is measured not by the average level, but by the standard deviation of the noise-only distribution. The important measure of a signal is the ratio of the signal strength to the standard deviation of the noise. The detector performance is characterized by this ratio and the threshold.

In most problems, a 10% false-alarm rate is too high. In standard statistical testing one often talks about "confidence" values of 5% or 1%. However, in most signal-processing applications the false-alarm rate must be several orders of magnitude lower for the system to be useful. That is because the rate is an individual-detection value. Consider, for example, a multibeam radar. Returns from a single pulse may come in from 100 directions. In each direction there may be on the order of 10,000 range bins. This means that on each pulse, there are on the order of 10^6 opportunities for a false alarm. In order to avoid overloading the operator, it may be necessary to keep the system false-alarm rate below about one per 100 pulses.

Figure 2 shows another way to analyze the performance of a detector. This is the same detector discussed in Fig. 1, but the signal strength is now treated as a variable. This means that for a given threshold setting the probability of detection depends on the signal strength. Figure 2 shows how the probability of detection varies with the signal-to-noise ratio. In this case, the threshold has been set for a false-alarm rate of 10^{-6}. Figure 2 illustrates an important point that is common with most detectors operating at these low false-alarm rates: The transition from a very low probability of detection to a very high probability of detection occurs over a fairly narrow decibel range. This is consistent with experience in auditory testing. Initially, the investigator sets the signal strength very low and the subject hears nothing. As the signal strength is increased, at some point the subject begins to hear the signal very faintly and with much uncertainty. By the time the signal strength has increased 3 dB beyond that point, the subject hears the signal clearly and calls it with no hesitation.

For this reason, there is usually no need to measure the probability of detection very accurately. Once the signal-to-noise ratio necessary for a 50% probability of detection is established, the 10% and 90% values are not far away.

The lower asymptote in Fig. 2 is not zero. It is the false-alarm rate. This is easy to see in Fig. 1, where the distributions become identical as the signal strength goes to zero.

LEAST-SQUARES PREDICTION AND ESTIMATION

Minimum-mean-squared-error estimation is among the most fruitful problems that have been investigated. Other criteria for goodness of fit have often been suggested, and in some cases may be more appropriate. However, they have not proven as rich in implications.

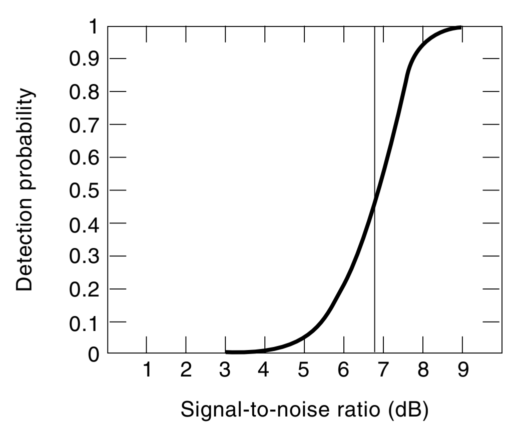

Figure 2. Probability of detection versus signal-to-noise ratio for the detector in Fig. 1. The false-alarm rate has been reduced to one per million. The transition from low probability of detection to high probability of detection occurs over a range of about 2 dB.

In the following discussion, the expectation of a variable, x, will be denoted by $\mathrm{E}[x]$, while the average of the available data will be denoted by $\langle x\rangle$. The transpose of a matrix or vector $\boldsymbol{x}$ will be denoted by $\boldsymbol{x}^{\mathrm{T}}$, while the complex conjugate of the transpose will be denoted by $\boldsymbol{x}^{\mathrm{H}}$. The trace of a matrix, A, will be denoted by tr A.

In its simplest form the problem is to estimate variable y from another variable, x. Since mean values are easy to add or remove, usually nothing is lost by assuming that both x and y are zero-mean. This means that the estimate $\hat{y}$ is equal to ax. The error criterion, or goodness-of-fit criterion, is $\epsilon = \mathrm{E}[(y - \hat{y})^2] = \mathrm{E}[y^2] - 2a\mathrm{E}[xy] + a^2\mathrm{E}[x^2]$, which is easily reduced to

$$\epsilon = \mathrm{E}[y^2]\left(1 - \frac{\mathrm{E}[xy]^2}{\mathrm{E}[x^2]\mathrm{E}[y^2]}\right) + \mathrm{E}[x^2]\left(a - \frac{\mathrm{E}[xy]}{\mathrm{E}[x^2]}\right)^2$$

This equation illustrates several common terms. Obviously the error ϵ is a minimum when the second term is zero, so the optimum coefficient is $a_{\mathrm{o}} = \mathrm{E}[xy]/\mathrm{E}[x^2]$. The quantity $\mathrm{E}[y^2]$ is called the *variance* of y, while the quantity $\mathrm{E}[xy]$ is called the *covariance* between x and y. The covariance can be normalized to give a *correlation coefficient* between x and y, $\mathrm{E}[xy]/\sqrt{\mathrm{E}[x^2]\mathrm{E}[y^2]}$, which is limited to the range between -1 and 1. The square of this quantity, $\mathrm{E}[xy]^2/(\mathrm{E}[x^2]\mathrm{E}[y^2])$, is called the *coherence* between x and y. The important point is that the coherence is the fraction of the variance of y that can be removed by the linear predictor. If the problem were turned around, so that y was used to predict x, the same coherence would still predict the fraction of the variance of x that could be removed.

This pattern of analysis also works when multiple variables are involved. In this case, it is convenient to group the coefficients and the independent variables, which may be complex, into column vectors, $\boldsymbol{a}$ and $\boldsymbol{x}$. The variable, y is then estimated by a scalar product $\hat{y} = \boldsymbol{a}^{\mathrm{H}}\boldsymbol{x}$. The error criterion is $\epsilon = \mathrm{E}[(y - \boldsymbol{a}^{\mathrm{H}}\boldsymbol{x})^*(y - \boldsymbol{a}^{\mathrm{H}}\boldsymbol{x})] = \mathrm{E}[y^*y] - \boldsymbol{a}^{\mathrm{H}}\mathrm{E}[y^*\boldsymbol{x}] - \mathrm{E}[y\boldsymbol{x}^{\mathrm{H}}]\boldsymbol{a} + \boldsymbol{a}^{\mathrm{H}}\mathrm{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{H}}]\boldsymbol{a}$. Again, it is convenient to define correlation quantities. The covariance matrix of $\boldsymbol{x}$ is $C_x = \mathrm{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{H}}]$. If $\boldsymbol{v} = \mathrm{E}[y^*\boldsymbol{x}]$ and $\sigma_y = \mathrm{E}[y^*y]$, then

$$\epsilon = \sigma_y - \boldsymbol{v}^{\mathrm{H}}C_x^{-1}\boldsymbol{v} + (\boldsymbol{a} - C_x^{-1}\boldsymbol{v})^{\mathrm{H}}C_x(\boldsymbol{a} - C_x^{-1}\boldsymbol{v})$$

Since a covariance matrix is necessarily positive definite, the last term is greater than or equal to zero, and can only be zero if $\boldsymbol{a}_{\mathrm{o}} = C_x^{-1}\boldsymbol{v}$, in which case the minimum mean squared error is $\epsilon_{\mathrm{o}} = \sigma - \boldsymbol{v}^{\mathrm{H}}C^{-1}\boldsymbol{v}$. This takes a more interesting form if one uses the total covariance matrix

$$C = \mathrm{E}\left[\begin{bmatrix} y \\ \boldsymbol{x} \end{bmatrix}\begin{bmatrix} y^* & \boldsymbol{x}^{\mathrm{H}} \end{bmatrix}\right] = \begin{bmatrix} \sigma_y & \boldsymbol{v}^{\mathrm{H}} \\ \boldsymbol{v} & C_x \end{bmatrix}$$

because the inverse is

$$C^{-1} = \begin{bmatrix} \dfrac{1}{\epsilon_{\mathrm{o}}} & -\dfrac{1}{\epsilon_{\mathrm{o}}}\boldsymbol{a}_{\mathrm{o}}^{\mathrm{H}} \\ -\dfrac{1}{\epsilon_{\mathrm{o}}}\boldsymbol{a}_{\mathrm{o}} & \left(C_{\boldsymbol{x}} - \dfrac{1}{\sigma_y}\boldsymbol{v}\boldsymbol{v}^{\mathrm{H}}\right)^{-1} \end{bmatrix}$$

By interchanging rows and columns it is easy to see that, in terms of the total covariance matrix and its inverse, it is arbitrary which element is being predicted. It follows that if a variable becomes extremely predictable, the corresponding diagonal element in the inverse will become very large. If any variable becomes completely predictable, the covariance matrix becomes singular.

This also provides a way to study multiple coherence. The multiple coherence of y with respect to a set of variables $x_1 \ldots x_n$, denoted $\mathrm{Coh}_{y;x_1,\ldots,x_n} = 1 - \epsilon_{\mathrm{o}}/\sigma_y$, is the fraction of the variance of σ_y that can be removed by a linear predictor based on the x's.

This forms the basis of some valuable methods for screening data. By computing the covariance matrix of experimental variables one can look for large correlations. This may be easier if the matrix is normalized so that the diagonal elements are all one, that is, $C_{i,j} \rightarrow C_{i,j}/\sqrt{C_{i,i}C_{j,j}}$. When the matrix is inverted, the diagonal elements will indicate if any variable is predicted especially well or poorly. If one diagonal element in the inverse is unusually small, it may be an indication that that variable somehow does not belong with the others. If one diagonal element of the inverse is unusually large, it can indicate that a variable is almost completely predicted from the others and therefore contributes little information to the data set. In the same vein, the optimum linear predictor for any variable can be extracted from the row or column of the inverse containing the corresponding diagonal element. The rows or columns of the inverse matrix can also be interpreted as a data-whitening filter.

The above pattern holds when the problem is generalized. A vector $\boldsymbol{y}$ can be estimated by a linear transformation $\hat{\boldsymbol{y}} = A^{\mathrm{H}}\boldsymbol{x}$. The important correlation matrices are $C_x = \mathrm{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{H}}]$, $C_y = \mathrm{E}[\boldsymbol{y}\boldsymbol{y}^{\mathrm{H}}]$, and $V = \mathrm{E}[\boldsymbol{x}\boldsymbol{y}^{\mathrm{H}}]$. In this case, the error quantity is $\epsilon = \mathrm{E}[(\boldsymbol{y} - \hat{\boldsymbol{y}})^{\mathrm{H}}(\boldsymbol{y} - \hat{\boldsymbol{y}})] = \mathrm{tr}\ \mathrm{E}[(\boldsymbol{y} - A^{\mathrm{H}}\boldsymbol{x})(y - A^{\mathrm{H}}x)^{\mathrm{H}}] = \mathrm{tr}[(A - C_x^{-1}V)^{\mathrm{H}}C_x(A - C_x^{-1}V)] + \mathrm{tr}(C_y - V^{\mathrm{H}}C_x^{-1}V)$. This, of course, immediately gives $\epsilon_{\mathrm{o}} = \mathrm{tr}(C_y - V^{\mathrm{H}}C_x^{-1}V)$ and $A_{\mathrm{o}} = C_x^{-1}V$. The total covariance matrix and its inverse take the form

$$\begin{bmatrix} C_y & V^{\mathrm{H}} \\ V & C_x \end{bmatrix}^{-1} = \begin{bmatrix} R_y & -(C_y^{-1}V^{\mathrm{H}})R_x \\ -A_{\mathrm{o}}R_y & R_x \end{bmatrix}$$

where $R_y = (C_y - V^{\mathrm{H}}C_x^{-1}V)^{-1}$ and $R_x = (C_x - VC_y^{-1}V^{\mathrm{H}})^{-1}$.

At first glance it may seem that the introduction of complex variables is an unnecessary complication. After all, the complex numbers can be treated as pairs of real numbers, so by doubling the size of the matrices we can solve the problem using only real variables. However, we cannot easily solve quite the same problem. The form $\hat{y} = ax$ carries an analytic assumption. For example, the solution can never take on the form Re y = Re x + Im x, Im y = Re x + Im x, because this would not be an analytic function. In order to make the problem equivalent one would have to pose the estimation as $\hat{y} = ax + bx^*$, giving up the analytic assumption. Usually the choice of complex variables for the original problem statement is determined by the physics of the problem. Thus, the use of complex variables injects certain *a priori* assumptions. If the complex functions seem appropriate to the problem definition, one should be careful about assuming that real variables could produce a sensible solution. Only if a nonanalytic solution can easily be given a physical interpretation should real variables be considered.

For the same reason, the most general form of the linear estimation procedure is rarely seen. It is $\hat{\boldsymbol{y}} = A\boldsymbol{x} + B\boldsymbol{x}^*$. However, even if an analytic solution is not necessary, it may be best to work the problem with complex variables in order to more easily give a physical interpretation to the problem definition or the solution.

MAXIMUM LIKELIHOOD, CRAMÉR–RAO, AND FISHER'S INFORMATION MATRIX

The predictors in the previous section are based on the constraint that a linear function is to be used. An obvious question is whether some nonlinear predictor could do better. As will be seen below, if the data are Gaussian, the answer is no. In non-Gaussian cases the minimum-mean-squared-error predictor is often difficult or impossible to find. However, even when this predictor cannot be found, it is sometimes possible to obtain bounds on how well any predictor can perform. Any candidate prediction function can then be compared with those bounds.

The most popular method to find such performance bounds is to use the Cramér–Rao inequality. The reasoning goes as follows.

An unknown quantity, A, is to be estimated. Here, A is a real number. Although A cannot be directly observed, an experiment is run that produces a real-number result, R, which depends in part on A. That is, the probability density function of R depends on A and can be written as $\text{prob}_{R|A}(r|A)$. The investigator must now make an estimate of A. The estimate, which depends on R, is denoted by $\hat{a}(R)$. The estimate may have a bias, $\beta(A) = \mathrm{E}[\hat{a}(R) - A]$. The question is "How good, in a mean-squared-error sense, can $\hat{a}(R)$ be?" The Schwarz inequality can be used to show that

$$\mathrm{E}\big[[\hat{a}(R) - A]^2\big] \geq \frac{\left(1 + \dfrac{d}{dA}\beta(A)\right)^2}{\mathrm{E}\left[\left(\dfrac{\partial}{\partial A}\ln \text{prob}_{R|A}(R|A)\right)^2\right]} = \frac{\left(\dfrac{d}{dA}\mathrm{E}[\hat{a}(R)]\right)^2}{\mathrm{E}\left[\left(\dfrac{\partial}{\partial A}\ln \text{prob}_{R|A}(R|A)\right)^2\right]}$$

or, equivalently,

$$\mathrm{E}\big[[\hat{a}(R) - A]^2\big] \geq \frac{-\left(\dfrac{d}{dA}\mathrm{E}[\hat{a}(R)]\right)^2}{\mathrm{E}\left[\dfrac{\partial^2}{\partial A^2}\ln \text{prob}_{R|A}(R|A)\right]}$$

This is of most interest when the estimator is unbiased, that is $\beta(A) = 0$. In this case the numerator of the right side of the above equations becomes one, and the right side of the equations is independent of the estimating procedure, $\hat{a}(R)$. Thus, for any unbiased estimator, one can arrive at bound on the goodness of the estimator in a mean-squared-error sense. Any estimator that gives equality with the bound is said to be *efficient.* No unbiased estimator can do better.

The above argument may give a good lower bound on the error, but it gives no help in finding a way to achieve that bound. One of the most intuitive lines of reasoning leads to the maximum likelihood estimator. It seems unreasonable to assume that the observation R was extremely improbable, given the true value of A. The extension of that idea is that the best guess for A is the one that would have made R seem most likely. The maximum likelihood estimate is the value of A that would maximize the probability density function $\text{prob}_{R|A}(R|A)$—in other words, the value of A that solves

$$\frac{d}{dA}\text{prob}_{R|A}(R|A) = 0$$

Often it is easier to solve

$$\frac{d}{dA}\ln \text{prob}_{R|A}(R|A) = 0$$

It turns out that if the maximum likelihood estimator exists, and if it is unbiased, then the maximum likelihood estimator is efficient.

Maximum likelihood estimators are often used with good results. For example, suppose that R is a Gaussian variable with unit variance and unknown mean. In order to estimate the mean from a single sample, consider that the logarithm of the probability density function is

$$\ln \text{prob}_{R|A}(R|A) = -\tfrac{1}{2}\ln(2\pi) - \tfrac{1}{2}(R - A)^2$$

so the maximum likelihood estimator is $\hat{A} = R$. The mean squared error is

$$-\left(\frac{\partial^2}{\partial A^2}\left[-\tfrac{1}{2}\ln 2\pi - \tfrac{1}{2}(R - A)^2\right]\right)^{-1} = 1$$

This is the best possible unbiased estimator in a mean-squared-error sense.

This idea can be generalized to the multivariable problem through the use of Fisher's information matrix. In this case a vector $\boldsymbol{A}$ is to be estimated after observing another vector $\boldsymbol{R}$ by use of an estimating function $\hat{\boldsymbol{a}}(\boldsymbol{R})$. The elements of Fisher's information matrix, J, can be defined in either of two equivalent ways:

$$J_{i,j} = \mathrm{E}\left[\frac{\partial \ln \text{prob}_{R|A}(R|A)}{\partial A_i}\frac{\partial \ln \text{prob}_{R|A}(R|A)}{\partial A_j}\right]$$

or

$$J_{i,j} = -\mathrm{E}\left[\frac{\partial^2 \ln \text{prob}_{R|A}(R|A)}{\partial A_i\, \partial A_j}\right]$$

To see how this works, consider the estimation error of the ith component of A. It has a bias error of $\beta_i(\boldsymbol{A}) = \mathrm{E}[\hat{a}_i(\boldsymbol{R}) - A_i]$ and a mean squared error of $\epsilon_i = \mathrm{E}[(\hat{a}_i(\boldsymbol{R}) - A_i)^2]$. It is convenient to define the vector $\boldsymbol{b}(i) = (\partial/\partial A_j)\mathrm{E}[\hat{a}_i(\boldsymbol{R})]$. (Of course, if the estimator is unbiased, then $\boldsymbol{b}(i)$ has a 1 in the ith position and zeros elsewhere.) Then

$$\epsilon_i \geq \boldsymbol{b}(i)^T J^{-1}\boldsymbol{b}(i)$$

It is important not to confuse the concept of efficiency with optimality. Arguments that an estimator is optimum must be based on game theory or decision theory. This mistake is tempting partly because it seems intuitively that an unbiased estimator should be better than a biased estimator. However, this is not necessarily true. The following problem illustrates the difficulty.

Suppose one needs to estimate the variance of a zero mean Gaussian variable.

$$\ln \text{prob}_{R|A}(R|A) = -\tfrac{1}{2}\ln 2\pi - \tfrac{1}{2}\ln A - \frac{R^2}{2A}$$

The Cramér–Rao bound is

$$\frac{-1}{\mathrm{E}\left[\dfrac{\partial^2}{\partial A^2}\left(-\tfrac{1}{2}\ln 2\pi - \tfrac{1}{2}\ln A - \dfrac{R^2}{2A}\right)\right]} = 2A^2$$

and the maximum likelihood estimator is $\hat{A} = R^2$. The following observations follow easily:

1. $\hat{A} = R^2$ is an unbiased maximum likelihood estimator of A.
2. $\hat{A} = R^2$ is an efficient estimator of A, that is, it meets the Cramér–Rao bound with equality.
3. $\hat{A} = R^2$ is obviously not an optimal, or even a good, estimator for A. In fact, if one were to ignore R and simply make $\hat{A} = 0$ the average mean squared error would only be half of that given by the efficient estimator.

In fact, a better estimator would be $\hat{A} = R^2/3$. It would have a mean squared error of only one-third that of the efficient estimator.

In many cases, especially those involving small sample sizes, it may be worthwhile to investigate the possibilities of biased estimators. Little seems to be known about how a good bias function can be chosen.

FOURIER TRANSFORMS AND SPECTRUM ESTIMATION

Fourier transforms can be viewed as the solution to a least-squared-error estimation problem. This is useful for analysis of existence, convergence, uniqueness, and so on. However, when designing analysis procedures it is much easier to think of them as a frequency translation and filtering process. Let $x(n)$ denote a sequence of data values sampled at regular intervals at a rate of f_s samples per second. Then $x(n)e^{-i2\pi nf/f_s}$ is a time sequence that has the same structure as $x(n)$ except that it is shifted so the information that was at frequency f is now at 0 Hz. The spectral coefficient for frequency f is now found by low-pass filtering with a summation filter function to get

$$\sum_{n=0}^{M-1} x(n)e^{-i2\pi nf/f_s}$$

For a time period $T = M/f_s$, the complex exponentials at $f = f_s m/M$ are uncorrelated, where m is any integer. So the Fourier transform components are defined as

$$X(m) = \eta \sum_{n=0}^{M-1} x(n)e^{-i2\pi mn/M}$$

This formula is often referred to as the discrete Fourier transform (DFT). Since the original data sequence can be recovered by

$$x(n) = \frac{1}{\eta M}\sum_{m=0}^{M-1} X(m)e^{i2\pi mn/M}$$

the transformation has lost no information.

The choice of η is arbitrary and usually depends on the software package used. Most standard programming packages use $\eta = 1$, and the reader can safely assume this for the following discussion. However, a few packages [e.g., MathCad (MathSoft, Inc.)] use $\eta = 1/\sqrt{M}$, which makes the above formulas symmetrical.

It is important to keep track of the exact form of the Fourier transform used, because it determines the form of Parseval's theorem. With the above definitions, Parseval's theorem says that

$$\sum_{n=0}^{M-1} x^*(n)x(n) = \frac{1}{\eta^2 M}\sum_{m=0}^{M-1} X^*(m)X(m)$$

This is the key to computing the power spectral density.

The *power spectral density* $S_P(m)$ of a waveform is a function that, when integrated over a frequency band, will give the power of the waveform in that band. The equations must be calibrated in order to make the integral over the total frequency band come out right. Since the frequency resolution of the analysis is f_s/M, the approximations to Riemann integrals look like

$$\text{power} = \frac{1}{M}\sum_{n=0}^{M-1} x^*(n)x(n) = \sum_{m=0}^{M-1} S_P(m)\frac{f_s}{M}$$

so Parseval's theorem gives

$$S_P(m) = \frac{1}{\eta^2 f_s M} X^*(m)X(m)$$

This gives a procedure for stationary waveforms. However, if one needs to analyze impulsive functions a different line of thought is necessary. An *impulse* will be defined here as a function that takes on nonzero values only within the interval $0 \le n < M$. In this case, power is not an interesting quantity and the energy in the waveform becomes important. The energy spectral density is found by first noting the energy in the waveform is

$$\text{energy} = \sum_{n=0}^{M-1} x^*(n)x(n)\frac{1}{f_s} = \sum_{m=0}^{M-1} S_E(m)\frac{f_s}{M}$$

In this case, Parseval's theorem gives

$$S_E(m) = \frac{1}{\eta^2 f_s^2} X^*(m)X(m)$$

In either case, when the results are plotted, the usual procedure is to plot the spectral density versus frequency on a decibel scale. If the original level of E[x^*x] was specified in decibels relative to a reference level, the spectral data should be plotted in decibels per hertz relative to the reference level. The spectrum should not be labeled as "per Hz$^{1/2}$" unless the author really intends that the function is to be integrated with respect to the square root of the frequency. This mistake is made by a remarkable number of authors.

When the waveform contains pure tonals (defined as signals whose bandwidth is less than the analysis resolution), special problems arise. A pure sinusoid would have an infinite power spectral density, and be properly modeled as a Dirac delta function in frequency. This cannot be sensibly plotted on the same scale as power that is distributed over an identifiable frequency band. In this case, the total power in the tonal should be estimated. Then the peak should be deleted from the plot, and replaced by a line indicating the sinusoidal power. For example, suppose the spectral density in the neighborhood of 60 Hz is 150 dB/Hz, while the indicated power spectral density for the 60 Hz bin is 160 dB/Hz. If the frequency resolution from the Fourier transform were 1/50 Hz, this would mean that the power in the line was 160 dB − 17 dB = 143 dB (since 10 log 1/50 = −17). When the data are reported, the plot should show a smooth spectrum at 150 dB/Hz through the 60 Hz region and a vertical line rising to a level of 143 dB.

The above formulas are usually considered to be a bad way to estimate a spectrum, because of sidelobe leakage. Therefore a window function, $w(n)$, is usually used. To see the effect, it is easiest to think of the equivalent low-pass filter.We can write the Fourier transform as a convolution filter:

$$X(f,n) = \sum_{L=0}^{M-1} w(L)x(n-L)e^{-i2\pi(n-L)f/f_s}$$

Then the Fourier transformation consists in sampling this function at regular intervals. The intervals are not necessarily simply related to the length of the Fourier transform.

In the first case, $w(n)$ was a *boxcar* filter. That is, $w(n) = 1$ if $0 \le n < M$, and $w(n) = 0$ otherwise. Many good window functions are known.

When analyzing a window, it helps to compare it with the boxcar window. Relative to a boxcar window, the more popular window functions widen the main-lobe frequency response, reducing the frequency resolution, in order to lower the sidelobes. A window is usually judged by two criteria: How much does it broaden the main lobe, and how much does it lower the sidelobes? The Dolph–Chebyshev window has the lowest worst-case sidelobe for a given main-lobe width. Although this window is rarely used, it is a quick way to see what can be done. The window shape, in the frequency domain, is controlled by a parameter β. For an M-point window,

$$W(f) = T_{M-1}\left(\frac{\cos(\pi f/f_s)}{\cos(\pi\beta/M)}\right)$$

where T_{M-1} is a Chebyshev polynomial of order $M - 1$. (This works because the Chebyshev polynomials are themselves solutions of a minimax problem.) The first zero of a boxcar window of the same length would be at f_s/M. The first zero of a Dolph–Chebyshev window is at approximately $\beta f_s/M$. If the window width were measured to the points 3 dB or 6 dB down, the width would be about $\sqrt{\beta}$ times the width of the boxcar window of the same length. The sidelobes would be about $27.3\beta - 6$ dB down from the main lobe. Thus, if a given level of sidelobe rejection is specified, one can immediately see how narrow a main lobe is possible (i.e., what the best possible frequency resolution is). Or if the frequency resolution of the window is specified, one can see how much sidelobe rejection is possible.

The Dolph–Chebyshev window is rarely used, for two reasons. First, although the worst sidelobes are well down, all of the other sidelobes are equally high. They do not taper off. Second, the endpoints of the window are often quite high. These problems are sometimes alleviated by convolving the Dolph–Chebyshev window with a short binomial window. The binomial window has a very broad main lobe but no sidelobes at all. When the two windows are convolved in the time domain, they are multiplied in the frequency domain. The time domain convolution smooths out the spikes at the end of the window, while the frequency domain multiplication reduces the distant sidelobes.

The Kaiser–Bessel window is a more popular choice. It is obtained by sampling the function

$$w(t) = \begin{cases} \dfrac{1}{T} I_0\left(\pi\beta\sqrt{4\left(\dfrac{t}{T}\right)\left(1-\dfrac{t}{T}\right)}\right), & 0 \le t \le T \\ 0 & \text{otherwise} \end{cases}$$

where T is the time duration of the window and I_0 is a Bessel function (1) computed by

$$I_0(z) = 1 + \frac{z^2/4}{(1!)^2} + \frac{(z^2/4)^2}{(2!)^2} + \frac{(z^2/4)^3}{(3!)^2} + \cdots$$

The first (and worst) sidelobe of the filter is approximately $27.3\beta - 20 \log \beta - 2.5$ dB down from the peak response. The first null occurs at approximately $\beta(1 + 0.333/\beta^2)/T$ Hz. If a boxcar filter is used, the first null occurs at a frequency of $1/T$ Hz.

The actual process of computing the Fourier transforms is usually done using an algorithm called the *fast Fourier transform* (FFT). It provides a much faster computation with less rounding error than one would get from a DFT. For present purposes it is important only to understand that the FFT has no theoretical significance. It is simply a quick way to compute the same result that could otherwise be obtained with a DFT.

To be efficient the FFT requires that M be a highly composite number, usually a power of 2. Since the length of the available data string is unlikely to be a power of 2, this might seem to be a problem. However, the problem is easily solved by padding the data with zeros to fill out the input vector. The effect of this is to overresolve the spectrum. This turns out to be very beneficial if the spectrum contains sharp features that might otherwise be difficult to resolve. Of course, one could get the same effect by interpolation, but it would be much more difficult. This works out so well that often the analyst may make M much longer than the size of the data set in order to get a smooth spectrum that is easy to interpret.

The use of a window and zero padding requires a modification of the equation calibration. Parseval's theorem again provides guidance.

STATIONARITY ISSUES

Most signal-processing theory assumes time-stationary processes, at least for the noise. All physical systems are ultimately not time-stationary. It is often important to arrive at some clear opinion about how nearly stationary the data are. A time sequence is time-stationary if, for any set of values $x(n)$, $x(n+1)$, . . ., $x(n+M-1)$ and any function $a(x(n)$, $x(n+1)$, . . ., $x(n+M-1))$ of those points, the average of a, or $E[a(x(n), x(n+1), \ldots, x(n+M-1))]$, is not a function of n. In other words, absolute time has no meaning for the sequence. This condition is usually impossible to test and stronger than is needed for most analyses. Therefore, it is much more common to assume that the data are *wide-sense stationary,* or *second-order stationary.* This means simply that all of the first and second moments of the data stream are independent of time. In this case it is possible to identify an autocorrelation function, $A(n) = E[x(t)x(t+n)]$, where $A(n)$ is independent of t.

When the data sequence is not second-order stationary, it is often possible to choose the Fourier transform lengths so that the spectrum changes slowly relative to the Fourier transform interval. Then sequential spectra can be compared and peaks in the spectrum followed as they change in time. When several spectra are plotted, one above another, on a single display, these peaks often follow characteristic paths down the display. Since the peaks trace out a visible line, narrowband components of a spectrum are often referred to as *lines.* Often a great deal of study and experience is required to interpret these lines. For example, lines at frequencies that are harmonics of power frequencies (50 Hz or 60 Hz, depending on the country) are apt to be a symptom of instrumentation problems.

Assuming that the data sequence is stationary, two assumptions are usually made that are only approximately true. The first is that Fourier coefficients corresponding to different frequencies are uncorrelated, that is, $E[X^*(m_1)X(m_2)] = 0$ for all $m_1 \neq m_2$. The second is that the real and imaginary parts of the Fourier coefficients are uncorrelated and of equal variance, that is, $E[X_r(m)X_i(m)] = 0$ and $E[X_r^2(m)] = E[X_i^2(m)]$ for all m. Another way to state the condition is that $E[X^2(m)] = 0$. As will be seen below, this second condition means that for Gaussian data the probability density function of $X(m)$ depends only on the magnitude of $X(m)$. Equal-probability contours of $X(m)$ then are circles in the complex plane, so the variables are called *circular.*

The Fourier coefficients from different frequency bins usually have a small but nonzero correlation because of sidelobe leakage in the window function. The amount of correlation is controlled by the choice of window and the extent to which the data have been whitened prior to study.

The circularity issue has not been fully explored. In order to do so, it is probably useful to define a *circularity anomaly,*

$$\alpha_c(m) = -\eta^2 \sum_{n-0}^{M-1} A(n) \sin \frac{2\pi nm}{M}$$

that is, the sine transform of the autocorrelation function. Then

$$E[X^2(m)] = \frac{2\alpha_c(m)}{\sin(2\pi m/M)} e^{i2\pi m/M}$$

The phase angle is independent of the spectrum. Under some circumstances, it is possible that this phase angle might be used as a test of stationarity. However, no such test procedures have been worked out.

If one suspects that circularity might not hold, it may be a good precaution to multiply each Fourier coefficient by $e^{-i2\pi m/M}$. This will have the effect of decorrelating the real and imaginary components. However, it will also maximize the difference in their magnitudes.

Usually $\alpha_c(m)$ is small enough to ignore safely. Therefore, the following sections will assume that the Fourier coefficients are circular. However, it is not clear when the rare exceptions occur. They are associated with steep changes in the spectrum. It is possible, in situations where a narrowband signal is on a steep spectral slope, that the signal will be more detectable on looking only at one part of the Fourier coefficients. This is not commonly done.

BANDWIDTH AND TIME–BANDWIDTH PRODUCTS

The entire frequency range available for analysis is usually wider than the signals of interest. Often, the signal energy is confined between two frequencies, f_1 and f_2. Then it is convenient to define a frequency bandwidth $W = f_2 - f_1$. Recalling that the integration time of the Fourier transform is $T = M/f_s$, the frequency resolution of the analysis is $1/T = f_s/M$, so there are $K = WT$ Fourier transform bins that contain the signal. K is referred to as the *time–bandwidth product* of the signal. It is often important to know what the duration and the bandwidth of the signal are.

Curiously, there is no generally agreed-upon way to define the bandwidth of a signal. Indeed, a similar problem may exist in defining the duration of a signal. Sometimes the nature of the problem may dictate a definition that is appropriate only to that problem. For example, when considering the uncertainty principle, the bandwidth and time duration of a signal are defined by

$$W^2 = \int_{-\infty}^{\infty} f^2 S(f)\,df \quad \text{and} \quad T^2 = \int_{-\infty}^{\infty} t^2 |x(t)|^2\,dt$$

when the signal is normalized so that

$$1 = \int_{-\infty}^{\infty} |x(t)|^2\,dt = \int_{-\infty}^{\infty} S(f)\,df$$

This leads to the uncertainty principle (2)

$$WT \geq 1/4\pi$$

However, these definitions for W and T seem not to be used for any other purpose than proving the uncertainty principle.

More frequently, the edges of the frequency band are defined as the points at which the spectrum is 3 dB down from the peak. This is especially appropriate when working with Butterworth filters. In this case the 3 dB down frequency is known as the *corner frequency.* This identity of the corner frequency and the 3 dB down frequency is not true for most other filter types, but they drop off fast enough that the error is small.

Square law detection theory can provide another useful definition of bandwidth. As will be seen below, in this case the detectability of a random signal increases as the square root of the time–bandwidth product and the average signal-to-noise ratio across the frequency band. Thus, for maximum detectability one would want to choose f_1 and f_2 to maximize

$$\frac{1}{\sqrt{f_2 - f_1}} \int_{f_1}^{f_2} \frac{S(f)}{N(f)} df$$

This prescription leads to

$$\frac{S(f_1)}{N(f_1)} = \frac{S(f_2)}{N(f_2)} = \frac{1}{2}\frac{1}{f_2 - f_1} \int_{f_1}^{f_2} \frac{S(f)}{N(f)} df$$

In other words, the edges of the frequency band should be chosen so that the signal-to-noise ratio at each edge is 3 dB below the average signal-to-noise ratio over the band.

SINGLE-WAVEFORM TESTING AND SQUARE LAW DETECTORS

One of the most instructive and fundamental detection problems involves a waveform that may or may not be present in Gaussian noise. The two possibilities are denoted as H_0 (no signal present) and H_1 (one signal present). The noise is assumed to be from a time-stationary random process, and is known only by its spectrum. This can be denoted by $\nu(m) = \mathrm{E}[X^*(m)X(m)|H_0]$.

The problem usually takes one of two different forms, depending on what *a priori* information about the signal is assumed. In the first case the signal is assumed to be known exactly. This is appropriate for study of active sonar or radar. The signal is then a waveform that takes on nonzero values only for a limited time. The Fourier transform of the signal will be denoted by $S(m)$, with the signal power designated as $\sigma(m) = S(m)^*S(m)$. Several lines of thought lead to use of a correlator or a convolution operator. This detector may be implemented in either the time domain or the frequency domain. However, it is easier to analyze in the frequency domain. The detector uses a linear filter described by $H(m)$ and is equivalent to forming a test statistic

$$U = \sum \mathrm{Re}[H(m)X(m)]$$

where the sum is taken over the $K = WT$ frequency bins that contain significant signal energy. The issue is the statistics of U,

$$\mathrm{E}[U|H_0] = 0 \quad \text{and} \quad \mathrm{E}[U|H_1] = \sum \mathrm{Re}[H(m)S(m)]$$

The only other quantity of interest is the variance,

$$\mathrm{E}[U^2|H_0] = \frac{1}{2} \sum H^*(m)H(m)\nu(m)$$

It can be easily shown that the optimum choice of filter function is $H(M) = S^*(m)/\nu(m)$. However, nothing is lost by scaling $H(M)$ so that $\mathrm{E}[U^2|H_0] = 1$. It also helps to designate the average signal-to-noise ratio over the band as $\langle\sigma/\nu\rangle_\mathrm{f}$. This is a different type of average than used above. It enables one to separate the effect of averaging, or time–bandwidth product, from the effect of the energy ratios. If this is done, then

$$\mathrm{E}[U|H_1] = \sqrt{2K\langle\sigma/\nu\rangle_\mathrm{f}}$$

If the noise is white (i.e., the spectrum is flat over the band) the above detector is called a *matched filter*. In this case, the signal-to-noise ratio reduces to the ratio of the total signal energy to the noise power spectral density.

The detection process consists in selecting a threshold value, U_th, and comparing it with the filter output. The false-alarm rate, or probability of false alarm, can be found from the usual Gaussian distribution

$$P_\mathrm{F} = Q(U_\mathrm{th})$$

where $Q(\ )$ is characterized by Eq. (26.2.3) of Ref. 1. Possibly useful values are

$$Q(3.72) = 10^{-4}$$
$$Q(4.75) = 10^{-6}$$
$$Q(5.61) = 10^{-8}$$
$$Q(6.36) = 10^{-10}$$

The probability of detection is at least 50% if $2K\langle\sigma/\nu\rangle_\mathrm{f} \geq U_\mathrm{th}^2$. If a = log U_th, then the signal excess can be defined as

$$\mathrm{SE_c} = 10\log\langle\sigma/\nu\rangle_\mathrm{f} + 10\log K + 3 - 2a$$

Two points may surprise the knowledgeable reader. First, SE is not a simple function of integrated signal power and integrated noise power. The averaging is done only after the ratio has been taken. The second is the 10 log dependence on K. This is an important difference between detection of a known signal and of an unknown signal (discussed below).

If the noise is white,

$$\mathrm{SE_c} = 10\log(\text{total signal power}) - 10\log(\text{noise power spectral density}) + 3 - 2a$$

The quantity $3 - 2a$ is sometimes referred to as the "recognition differential." However, this term is used in a confusing variety of different ways, so one is usually better off to avoid using it altogether.

The second important variant on this problem is the unknown signal. In this case, the signal is assumed to be a time-stationary Gaussian signal known only by its spectrum, which will be denoted by $\sigma(m) = \mathrm{E}[X^*(m)X(m)|H_1] - \mathrm{E}[X^*(m)X(m)|H_0]$. Several arguments lead to a square law detector,

$$V = \sum X^*(m)X(m)H(m)$$

Assuming the signal spectrum is accurately known, the best choice of the frequency weights is

$$H(m) = \frac{\sigma(m)}{\nu(m)[\nu(m) + \sigma(m)]}$$

This is approximated by the Eckart filter, $H(m) = \sigma(m)/\nu^2(m)$ (3). However, for various reasons, including difficulty in knowing the signal spectrum, it is often more practical to use a noise-whitening filter followed by a band-pass filter,

$$H(m) = \frac{1}{K\nu(m)}$$

This means that

$$\mathrm{E}[V|H_0] = 1 \quad \text{and} \quad \mathrm{E}[V|H_1] = 1 + \left\langle \frac{\sigma}{\nu} \right\rangle_{\mathrm{f}}$$

At this point it is tempting to use the central limit theorem (CLT) to argue that the distribution of V is Gaussian and the detection statistics can be estimated as above. However, the CLT works poorly on the tails of the distribution, and fortunately this approximation is not necessary. In fact, V has the form of a chi-square variable, and the distribution of V is the gamma distribution. If $\Gamma(\ ,\ )$ denotes the incomplete gamma function, then the probability density function of V is

$$G(K, KV) = \frac{\Gamma(K, KV)}{\Gamma(K)}$$

This equation differs slightly from Eq. (26.4.19) of Ref. 1 because of different normalization and because K is only half the number of degrees of freedom.

Then if V_{th} denotes the threshold,

$$P_{\mathrm{F}} = G(K, KV_{\mathrm{th}})$$

Exact evaluation of this equation is cumbersome. However, it is easily approximated by

$$\frac{(KV)^{K-1}}{(K-1)!} e^{-KV} < G(K, KV) < \frac{1}{1 - \dfrac{K-1}{KV}} \frac{(KV)^{K-1}}{(K-1)!} e^{-KV}$$

In fact, $G(K, KV)$ stays much closer to the upper bound.

This gives an easy way to estimate P_{F} for various values of K and V_{th}. However, it is also useful to be able to turn the problem around and find the required V_{th} for a given P_{F} and K. In most cases this problem cannot be solved in closed form. It has been found empirically that, for realistic false-alarm rates, a good approximating equation is

$$10 \log(V_{\mathrm{th}} - 1) = a - 5 \log K + 10 \log \left(1 + \frac{b}{\sqrt{K}}\right)$$

For this purpose, the following table may be adequate:

$$\begin{aligned} P_F &= 10^{-4}, & a &= 5.705, & b &= 1.2 \\ P_F &= 10^{-6}, & a &= 6.77, & b &= 1.65 \\ P_F &= 10^{-8}, & a &= 7.49, & b &= 2 \\ P_F &= 10^{-10}, & a &= 8.03, & b &= 2.4 \end{aligned}$$

As above, one can define a signal excess equation as

$$\mathrm{SE_s} = 10 \log \left\langle \frac{\sigma}{\nu} \right\rangle_{\mathrm{f}} + 5 \log WT - a - 10 \log \left(1 + \frac{b}{\sqrt{K}}\right)$$

The last term may be interpreted as the error that would have resulted if a Gaussian distribution assumption had been made for V.

For detection of tonals, this equation takes a somewhat different form because a different definition of σ is used. When investigating tonals, or nearly pure sinusoids, instead of specifying the power spectral density of the signal, only the total signal power is specified. The spectrum of the signal is then treated as a Dirac delta function times that signal power. The key assumption is that the total width of the signal is less than the width of a Fourier bin. Then the apparent signal power spectral density depends on the bin width, which is now W. With this new different definition of σ,

$$\mathrm{SE_s} = 10 \log \left\langle \frac{\sigma}{\nu} \right\rangle + 5 \log T - 5 \log W - a - 10 \log \left(1 + \frac{b}{\sqrt{K}}\right)$$

As suggested above, it is often difficult to obtain good *a priori* information about the signal. However, similar problems occur for the noise. If the absolute level of the noise is unknown, then it must be measured before thresholds can be set. When attempting to detect narrowband signals, this process is referred to as *noise spectrum equalization,* or NSE. In its simplest form, NSE can be analyzed as follows.

When attempting to detect a narrowband signal, a common approach is to plot the power spectral density and look for sharp narrow peaks. The eye can then easily identify the average level of the noise and judge whether a particular spike is significantly higher than that average level. To quantify this, assume that L bin levels around the signal bin are averaged. If K is the time–bandwidth product for each bin, then the average noise level is being estimated with a time–bandwidth product of KL. What the eye actually sees, especially if the spectrum is plotted on a decibel scale, is the ratio of the estimated power in the signal bin to the estimated power in the surrounding noise bins. This is a ratio of two powerlike variables. This ratio will have an F distribution. Let $\rho = \sigma(m)/\nu(m)$ denote the signal-to-noise ratio in the signal bin, and assume that the noise spectrum is flat over the L frequency bins around the signal. The probability density function of the ratio is

$$\mathrm{prob}_Z(z) = \frac{(K + KL - 1)!(KL)^{KL}}{(K-1)!(KL-1)!} \times \left(\frac{K}{1+\rho}\right)^K \frac{z^{K-1}}{\left(\dfrac{zK}{1+\rho} + KL\right)^{K+KL}}$$

The cumulative probability function can be written as

$$\mathrm{Prob}_Z(z) = \frac{B_{z/z+L(1+\rho)}(K, KL)}{B(K, KL)} = I_{z/(z+L(1+\rho))}(K, KL)$$

where $B_\xi(K, KL)$ is the incomplete beta function (1). To compute the false-alarm rate, simply set $\rho = 0$.

This type of detector can work very well when the time–bandwidth products are large. However, for small time–bandwidth products the price of having to estimate the normalization factor is severe. The extreme case occurs when $K = L = 1$. In this case, the false-alarm rate, for a threshold value of th, is $1/(\mathrm{th} + 1)$. This means that if one wanted a false-alarm rate of 10^{-4}, it would take a signal-to-noise ratio of 40 dB to give a 50% probability of detection. As the time–bandwidth product of the detector increases, the detection performance improves rapidly, approaching the performance of a square law detector as L becomes large. This pattern—the importance of large time–bandwidth product when a normalization factor is estimated from the data—will reappear below in the discussion of two-channel detectors.

The general problem of detection of Gaussian signals in Gaussian noise, or even of sinusoids in Gaussian noise, is far from solved. For example, effects of sidelobe leakage in the Fourier transforms have been ignored. More importantly, if the noise power spectral density is significantly far from white, the resulting detection statistic is a sum of unequal chi-square variables. The probability distribution of such a variable is so cumbersome as to be nearly useless. A good method for approximating it is needed.

TWO-CHANNEL DETECTION

Detection or estimation of a signal that is believed to be common to two different waveforms may be done in several different ways, depending on the *a priori* information available and the type of information to be extracted. In the following discussion, $X(n)$ and $Y(n)$ are the two complex data sequences. Usually they are Fourier coefficients from successive transform intervals. In the following equations, $\langle\ \rangle$ denotes the average over K data samples. It will also be initially assumed that the signal-to-noise ratio in both sequences is the same. That is, $E[X^*X|H_0] = E[Y^*Y|H_0] = \nu$ and $E[X^*X|H_1] = E[Y^*Y|H_1] = \nu + \sigma$. The noise will be assumed to be Gaussian, uncorrelated between the two sequences, and independent of the signal.

There are four principal functions from which one may choose:

$$\begin{aligned}
u_1 &= \langle X^*X\rangle \overset{?}{>} T_1\nu && \text{square law}\\
u_2 &= \mathrm{Re}\langle X^*Y\rangle) \overset{?}{>} T_2\nu && \text{correlator}\\
u_3 &= \frac{\mathrm{Re}\langle X^*Y\rangle}{\sqrt{\langle X^*X\rangle\langle Y^*Y\rangle}} \overset{?}{>} T_3 && \text{correlation coefficient}\\
u_4 &= \frac{|\langle X^*Y\rangle|^2}{\langle X^*X\rangle\langle Y^*Y\rangle} \overset{?}{>} T_4 && \text{coherence}
\end{aligned}$$

The first function is included as a reference. It is the simple square law detector, analyzed previously. It forms a baseline for judgement of the other detectors, since it simply uses one of the two sequences. The comparison gives an indication of the value of having two sequences instead of one. An important case that is not considered here is $\langle(X + Y)^*(X + Y)\rangle$. This is because it does not really constitute a separate case. It is simply the square law detector with a 3 dB increase in signal-to-noise ratio.

In each case, the quantity u is compared with a threshold. (In the first two cases it is necessary to know ν in order to set the threshold.) It is important to know how the false-alarm rate will be determined by the threshold. However, this is only part of the story, since the probability of detection is also important. In each case, it is possible to associate a signal-to-noise ratio with the threshold that will produce approximately a 50% probability of detection. The critical signal-to-noise ratios are

$$\begin{aligned}
1+\left(\frac{\sigma}{\nu}\right)_1 &= T_1\\
\left(\frac{\sigma}{\nu}\right)_2 &= T_2\\
\frac{(\sigma/\nu)_3}{(\sigma/\nu)_3+1} &= T_3, \quad \text{or} \quad \left(\frac{\sigma}{\nu}\right)_3 = \frac{T_3}{1-T_3}\\
\frac{(\sigma/\nu)_4^2}{[(\sigma/\nu)_4+1]^2} &= T_4, \quad \text{or} \quad \left(\frac{\sigma}{\nu}\right)_4 = \frac{\sqrt{T_4}}{1-\sqrt{T_4}}
\end{aligned}$$

These four signal-to-noise ratios will be called threshold signal-to-noise ratios. However, they are actually bin signal-to-noise ratios. To reconcile the following discussion with standard detection equations, one would have to at least correct for the bandwidth of the frequency bins. Further, due to asymmetries in the distribution functions, the threshold signal-to-noise ratios do not correspond precisely with a 50% probability of detection. The errors from this asymmetry are usually very small.

The false-alarm rates are, of course, determined by the threshold values. For the correlator the false-alarm rate is

$$P_F(T_2) = \frac{1}{2^{2K-1}(K-1)!}\sum_{n=1}^{K-1}\frac{(2K-n-2)!2^n}{n!(K-n-1)!}\Gamma(n+1, 2KT_2)$$

For the correlation coefficient detector the false-alarm rate is (4)

$$P_F(T_3) = \frac{(K-1)!}{\sqrt{\pi}\left(\dfrac{2K-3}{2}\right)!}\int_{T_3}^{1}(1-t^2)^{(2\kappa-3)/2}\,dt$$

Although this formula can be integrated in closed form, the solution is very cumbersome. However, it lends itself to numerical integration. For the coherence detector the false-alarm rate is (5)

$$P_F(T_4) = (1-T_4)^{K-1}$$

These formulas were used to compute Figs. 3, 4, 5, and 6. In each case the plot was designed to answer the question, "If

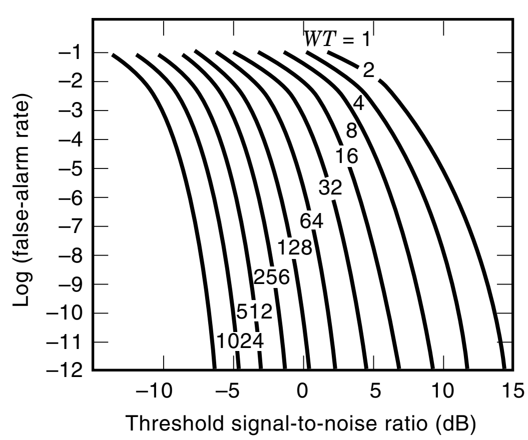

Figure 3. Log of false-alarm rate versus threshold signal-to-noise ratio for a square law detector. The curves are separated by about 1.5 dB for large WT products, but performance deteriorates more rapidly for small WT products.

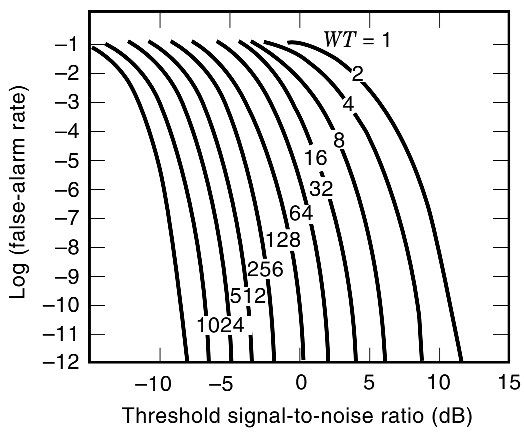

Figure 4. Log of false-alarm rate versus threshold signal-to-noise ratio for a correlator. The curves approximate those in Fig. 3 with a doubling of the *WT* product.

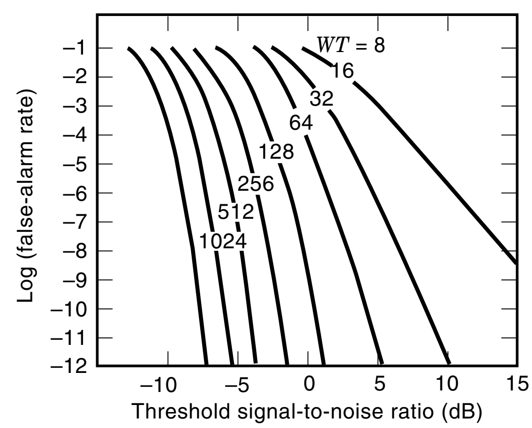

Figure 6. Log of false-alarm rate versus threshold signal-to-noise ratio for a coherence. Again, the performance deteriorates rapidly for small *WT* products.

the detector is set up to detect a signal at a given signal-to-noise ratio, what will the false-alarm rate of the detector be?" In each case, a threshold signal-to-noise ratio was chosen and the corresponding threshold value calculated. Then the probability of a noise-only false alarm was calculated and plotted. This was done for several values of $K = WT$, the time–bandwidth product. Since in general low false-alarm rates are necessary, the curves are mainly useful for the region of $P_{fa} < 10^{-4}$. The following discussion will address only this region.

In Fig. 3, for a given false-alarm probability, the curves are separated by about 1.5 dB in the large-*WT* cases. This agrees with the general rule that the integration gain of a detector is 5 log *WT*. However, for small *WT* values the separation increases to about 2.5 dB. This is because the 5 log *WT* is based on application of the CLT, which breaks down for small *WT*. In some cases this can lead to a difference of 3 or 4 dB in minimum detectable signal.

The curves in Fig. 4 nearly overlie those in Fig. 3, with a shift in *WT*. For example, the curve for $WT = 1$ in Fig. 4 closely overlays the curve for $WT = 2$ in Fig. 3. In other words, the advantage in having a second waveform and using a correlator over using a square law detector on one waveform is a factor of 2 in the integration time needed.

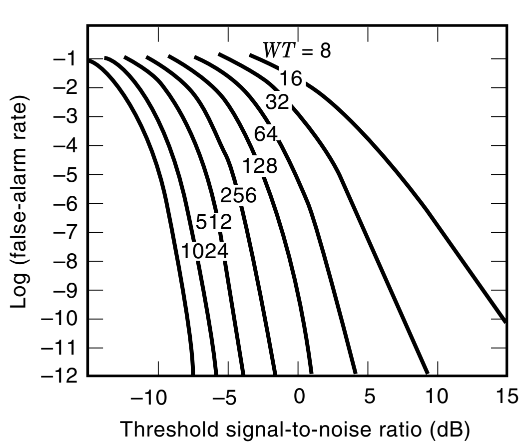

Figure 5. Log of false-alarm rate versus threshold signal-to-noise ratio for a correlation coefficient. The curves approximate those in Fig. 4 for large *WT* products but deteriorate rapidly for small *WT* products. This illustrates the difficulty of estimating a normalization factor from local data unless the *WT* factor is very large.

For large *WT*, the curves in Figs. 4, 5, and 6 nearly coincide. In other words, for large *WT*, all three of these techniques give nearly the same performance. Selection among these formulas can be made on the basis of considerations other than detection performance, such as ease of implementation.

For small *WT*, the performance of the normalized detectors deteriorates so rapidly that curves for *WT* less than 8 were not even plotted. This is consistent with the previous observations about normalized spectra. Normalized detection formulas work well only with large sample sizes. For *WT* less than 128, the normalized formulas do not work as well as a square law detector using only one sequence.

GAUSSIAN DISTRIBUTIONS

Most theoretical work on signal-processing problems assumes a Gaussian noise distribution. This assumption rests on two points of practical experience. First, much of the noise encountered in operating systems is approximately Gaussian. Second, data-processing systems based on Gaussian noise assumptions have a good track record in a wide range of problems. (This record is partly due to the coincidence between solutions based on Gaussian noise theory and solutions based on least-squares theory, as will be seen below.)

From a theoretical viewpoint the key feature of the Gaussian distribution is that a sum of Gaussian variables has a Gaussian distribution. (Other distributions with this property, called *alpha stability,* exist. One example is the Cauchy distribution. However, their role has yet to be established.) The importance of this fact is difficult to exaggerate. It means, among other things, that when Gaussian noise is passed through a linear filter, the output will still be Gaussian. (Unfortunately, little is known about what happens to the distribution of non-Gaussian noise when it is filtered. It is often said that because of the CLT the output of the filter can be assumed to be Gaussian. However, many important counterexamples are known, e.g., AM radio.) Partly because of this, the Gaussian distribution is almost the only distribution for which the extension to multiple variables or complex variables is understood.

The CLT is often cited as another reason to assume a Gaussian distribution. The CLT says that if a variable y is an average of a large number of variables, $x_1, x_2, \ldots, x_N$, then the distribution of y is approximately Gaussian and that this approximation improves as N increases, that is, y is asymptotically Gaussian. The necessary and sufficient conditions for this theorem are not known. However, several sets of sufficient conditions are known, and they seem to cover most reasonable situations. For example, one set of sufficient conditions is that the x_i's are independent and have equal variance.

The reader should, however, use some caution in invoking the CLT. It is an asymptotic result that is only approximately true for finite N. Further, the accuracy of this approximation is often very difficult to test. It tends to come into play fairly quickly in the central portions of the distribution, so when the experimental distribution is plotted the data look deceptively close to a Gaussian curve. However, detection and estimation problems tend to depend on the tails of the distribution, which may be very slow to converge to a Gaussian limit and cause large errors that are poorly understood. The investigator should always be alert for the possibility that a Gaussian distribution is not appropriate and should therefore consider alternatives.

Let $\boldsymbol{x}$ denote a column vector of real variables, $\boldsymbol{x}^T = [x_1 \; x_2 \; \cdots \; x_n]$, and let $C = \mathrm{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}]$ denote the covariance matrix of $\boldsymbol{x}$. Then the statement that $\boldsymbol{x}$ is Gaussian means that

$$\mathrm{prob}_{\boldsymbol{x}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n |C|}} e^{-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}} C^{-1} \boldsymbol{x}}$$

If the variables are complex, it is possible to define two important square matrices, $\Gamma = \langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{H}} \rangle$ and $C = \langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \rangle$. It is customary to assume that $C = 0$, which is the circularity assumption. This custom will be adopted later. In this case, it is convenient to define the accent vector

$$\acute{\boldsymbol{x}} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{x}^* \end{bmatrix}$$

The moment matrices and their inverses take the form

$$\mathrm{E}[\acute{\boldsymbol{x}}\acute{\boldsymbol{x}}^{\mathrm{H}}] = E = \begin{bmatrix} \Gamma & C \\ C^* & \Gamma^* \end{bmatrix} = \begin{bmatrix} A & B \\ B^* & A^* \end{bmatrix}^{-1}$$

The probability density function of $\acute{\boldsymbol{x}}$ is

$$\mathrm{prob}(\boldsymbol{x}) = \frac{1}{\pi^n \sqrt{|E|}} e^{-\boldsymbol{x}^{\mathrm{H}} A \boldsymbol{x} - (\boldsymbol{x}^{\mathrm{H}} B \boldsymbol{x}^* + \boldsymbol{x}^{\mathrm{T}} B^* \boldsymbol{x})/2}$$

For a single complex Gaussian variable x, this simplifies. Let $\gamma = \mathrm{E}[xx^*]$, let $c = \mathrm{E}[x^2]$, and let $\rho = c/\gamma$. Then

$$\mathrm{prob}(x) = \frac{1}{\pi\gamma\sqrt{1-\rho^*\rho}} \exp\left(\frac{-[x^*x - \frac{1}{2}(x^2\rho^* + z^{*2}\rho)]}{\gamma(1-\rho^*\rho)}\right)$$

Using the accent notation for the variables, the joint and conditional distributions take simple forms. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are jointly Gaussian vectors of length n and m respectively, the total covariance matrix can be defined as

$$E_{\mathrm{total}} = \mathrm{E}\left[\begin{bmatrix} \acute{\boldsymbol{x}} \\ \acute{\boldsymbol{y}} \end{bmatrix} [\acute{\boldsymbol{x}}^{\mathrm{H}} \quad \acute{\boldsymbol{y}}^{\mathrm{H}}]\right] = \begin{bmatrix} E_{xx} & E_{xy} \\ E_{yx} & E_{yy} \end{bmatrix} = \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}^{-1}$$

Then the joint distribution of $\boldsymbol{x}$ and $\boldsymbol{y}$ is

$$\frac{1}{\pi^{n+m}\sqrt{|E_{\mathrm{total}}|}} \exp\left(-\frac{1}{2}[\acute{\boldsymbol{x}}^{\mathrm{H}} \quad \acute{\boldsymbol{y}}^{\mathrm{H}}] E_{\mathrm{total}}^{-1} \begin{bmatrix} \acute{\boldsymbol{x}} \\ \acute{\boldsymbol{y}} \end{bmatrix}\right)$$

while the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{y}$ is

$$\mathrm{prob}(\boldsymbol{x}|\boldsymbol{y}) = \frac{\sqrt{|F_{11}|}}{\pi^n} e^{-\frac{1}{2}(\acute{\boldsymbol{x}} - E_{xy}E_{yy}^{-1}\acute{\boldsymbol{y}})^{\mathrm{H}} F_{11} (\acute{\boldsymbol{x}} - E_{xy}E_{yy}^{-1}\acute{\boldsymbol{y}})}$$

The moment generating function of a complex vector $\boldsymbol{s}$ is

$$\mathrm{mgf}(\boldsymbol{s}) \equiv \mathrm{E}e^{-\boldsymbol{s}^{\mathrm{H}}\boldsymbol{x} - \boldsymbol{x}^{\mathrm{H}}\boldsymbol{s}} = e^{\frac{1}{2}\acute{\boldsymbol{s}}^{\mathrm{H}} E \acute{\boldsymbol{s}}} = e^{\boldsymbol{s}^{\mathrm{H}}\Gamma\boldsymbol{s} + (\boldsymbol{s}^{\mathrm{H}} C \boldsymbol{s}^* + \boldsymbol{s}^{\mathrm{T}} C^* \boldsymbol{s})/2}$$

For a single variable, this simplifies to

$$\mathrm{E}e^{-s^*x + x^*s} = e^{s^*\gamma + (s^{*2}c + s^2c^*)/2}$$

Matching up coefficients for the fourth moments gives a little-known result,

$$\mathrm{E}[(x^*x)^2] = \gamma^2(2 + \rho^*\rho)$$

In other words, the *kurtosis,* defined here as the ratio of the fourth moment to the square of the second moment, varies between 2 and 3 depending on the degree of circularity of the variable. For real variables, $\rho = 1$, so the kurtosis is 3. For circular Gaussian variables, the most commonly used complex distribution, $\rho = 0$, so the kurtosis is 2. (Some authors subtract 3 from the ratio in their definition of kurtosis, so that for real Gaussian variables the kurtosis is zero. For formulations that include complex variables this is not a simplification.)

LIKELIHOOD DETECTORS FOR GAUSSIAN NOISE

Assuming a known signal, $\boldsymbol{s}$, in Gaussian noise the likelihood ratio for a sample variable $\boldsymbol{x}$ is

$$\frac{\dfrac{1}{\pi^n\sqrt{|E|}} e^{-\frac{1}{2}(\acute{\boldsymbol{x}} - \acute{\boldsymbol{s}})^{\mathrm{H}} E^{-1} (\acute{\boldsymbol{x}} - \acute{\boldsymbol{s}})}}{\dfrac{1}{\pi^n\sqrt{|E|}} e^{-\frac{1}{2}\acute{\boldsymbol{x}}^{\mathrm{H}} E^{-1} \acute{\boldsymbol{x}}}}$$

Isolating the terms that depend on $\boldsymbol{x}$, the likelihood ratio depends only on the expression

$$\acute{\boldsymbol{x}}^{\mathrm{H}} E^{-1} \acute{\boldsymbol{s}}$$

This provides a justification for the correlation structure discussed above.

The Gaussian signal assumption leads to a more complicated structure. In its simplest form, the signal is modeled as a random complex amplitude times a signal model vector $\boldsymbol{v}$

that is normalized so that $\boldsymbol{v}^{\mathrm{H}}\boldsymbol{v} = n$. If we admit that the signal may be noncircular, the signal covariance matrix takes a rank-two form:

$$P = \begin{bmatrix} \sigma \boldsymbol{v}\boldsymbol{v}^{\mathrm{H}} & c\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}} \\ c^*\boldsymbol{v}^*\boldsymbol{v}^{\mathrm{H}} & \sigma \boldsymbol{v}^*\boldsymbol{v}^{\mathrm{T}} \end{bmatrix}$$

$$= \begin{bmatrix} \sqrt{c}\boldsymbol{v} & \sqrt{c}\boldsymbol{v} \\ \sqrt{c^*}\boldsymbol{v}^* & -\sqrt{c^*}\boldsymbol{v}^* \end{bmatrix}$$

$$\begin{bmatrix} \dfrac{\sigma/\sqrt{c^*c}+1}{2} & 0 \\ 0 & \dfrac{\sigma/\sqrt{c^*c}-1}{2} \end{bmatrix} \begin{bmatrix} \sqrt{c^*}\boldsymbol{v}^{\mathrm{H}} & \sqrt{c}\boldsymbol{v}^{\mathrm{T}} \\ \sqrt{c^*}\boldsymbol{v}^{\mathrm{H}} & -\sqrt{c}\boldsymbol{v}^{\mathrm{T}} \end{bmatrix}$$

This notation can be simplified by introducing matrices V and D so that the above equation becomes $P = VDV^{\mathrm{H}}$. Ignoring terms that are independent of $\boldsymbol{x}$, the log of the likelihood ratio becomes

$$\acute{\boldsymbol{x}}^{\mathrm{H}}E^{-1}\acute{\boldsymbol{x}} - \acute{\boldsymbol{x}}^{\mathrm{H}}(E + VDV^{\mathrm{H}})^{-1}\acute{\boldsymbol{x}}$$

This simplifies to a quadratic form

$$\acute{\boldsymbol{x}}^{\mathrm{H}}E^{-1}VTV^{\mathrm{H}}E^{-1}\acute{\boldsymbol{x}} = \acute{\boldsymbol{x}}^{\mathrm{H}}W\acute{\boldsymbol{x}}$$

where T is a 2×2 matrix defined by

$$T^{-1} = D^{-1} + V^{\mathrm{H}}E^{-1}V$$

and W is a $2n \times 2n$ nonnegative matrix of rank 2. This provides justification for the square law detector discussed above.

OTHER DISTRIBUTIONS

As signal-processing applications become more sophisticated, other functions of complex variables come into play. For example, in the above discussions products of complex variables have already been encountered. In some deconvolution problems, quotients also arise.

The extension of standard probability theory to complex variables is an interesting exercise. The reason is that probability density functions are not analytic functions. (Obviously, they cannot be, since they always take on only real values.) Thus, standard theory of analytic continuation is not helpful. It seems that the easiest way to deal with this is simply to modify the basic definitions to accommodate the complex numbers and then do a set of derivations that parallel those already familiar for real variables. The following table shows some of the parallel formulas. (In the Gaussian case, only cir-

Real Variables	Complex Variables
	$dA_x = dx_r\, dx_i$
Probability density function	
$P_X(X < x) = \int_{-\infty}^{x} p_X(t)\, dt$	$P_X(X \in A) = \int\int_A p_X(x)\, dA_x$
Average	
$\mathscr{E}[X] = \int_{-\infty}^{\infty} x\, p_X(x)\, dx$	$\mathrm{E}[X] = \int\int_\infty x p_X(x)\, dA_x$
Gaussian	
$p_X(x) = \dfrac{1}{\sqrt{2\pi\nu}} e^{-x^2/2\nu}$	$p_X(x) = \dfrac{1}{\pi\nu} e^{-x^*x/\nu}$
Gaussian (multivariable)	
$p_{\boldsymbol{X}}(x) = \dfrac{1}{(2\pi)^{n/2}\sqrt{\lvert C \rvert}} e^{-1/2\boldsymbol{x}^{\mathrm{T}}C^{-1}\boldsymbol{x}}$	$p_{\boldsymbol{X}}(x) = \dfrac{1}{\pi^n \lvert C \rvert} e^{-\boldsymbol{x}^{\mathrm{H}}C^{-1}\boldsymbol{x}}$
Sum $Z = X + Y$	
$p_Z(z) = \int_{-\infty}^{\infty} p_{X,Y}(z - y, y)\, dy$	$p_Z(z) = \int\int_\infty p_{X,Y}(z - y, y)\, dA_y$
Product (general) $Z = XY^*$	
$p_Z(z) = \int_{-\infty}^{\infty} \dfrac{p_{X,Y}(z/y, y)}{\lvert y \rvert} dy$	$p_Z(z) = \int\int_\infty \dfrac{p_{X,Y}(z/y^*, y)}{y^*y} dA_y$
Product (Gaussian) $Z = XY^*$	
$p_Z(z) = \dfrac{1}{\pi\sqrt{1-\rho}} \exp\left(\dfrac{\rho z}{1-\rho^2}\right) K_0\left(\dfrac{z}{1-\rho^2}\right)$	$p_Z(z) = \dfrac{2}{\pi(1-\rho^*\rho)} \exp\left(\dfrac{\rho^* z + \rho z^*}{1-\rho^*\rho}\right) K_0\left(\dfrac{2\sqrt{z^*z}}{1-\rho^*\rho}\right)$
Quotient (general) $Z = X/Y$	
$p_Z(z) = \int_{-\infty}^{\infty} \lvert y \rvert\, p_{X,Y}(zy, y)\, dy$	$p_Z(z) = \int\int_\infty y^*y\, p_{X,Y}(zy, y)\, dA_y$
Quotient (Gaussian) $Z = X/Y$	
$p_Z(z) = \dfrac{\sqrt{1-\rho^2}}{\pi[(z-\rho)^2 + (1-\rho^2)]}$	$p_Z(z) = \dfrac{(1-\rho^*\rho)}{\pi[(z-\rho)^*(z-\rho) + (1-\rho^*\rho)]^2}$
Moment generating function	
$m_X(s) = \int_{-\infty}^{\infty} e^{-sx} p_X(x)\, dx$	$m_X(s) = \int\int_\infty e^{-(s^*x + x^*s)} p_X(x)\, dA_x$
Gaussian moments	
$\mathrm{E}[X^{2i}] = \dfrac{(2i)!\,\nu^i}{2^i(i)!}$	$\mathrm{E}[(X^*X)^i] = i!\ \nu^i$
Fourth moment (Gaussian)	
$\mathrm{E}[X_1X_2X_3X_4] = \mathrm{E}[X_1X_2]\mathrm{E}[X_3X_4] + \mathrm{E}[X_1X_4]\mathrm{E}[X_3X_2] + \mathrm{E}[X_1X_3]\mathrm{E}[X_2X_4]$	$\mathrm{E}[X_1X_2^*X_3X_4^*] = \mathrm{E}[X_1X_2^*]\mathrm{E}[X_3X_4^*] + \mathrm{E}[X_1X_4^*]\mathrm{E}[X_3X_2^*]$

cular variables are considered.) For the real variable case, $\rho = E[xy]$. For the complex case $\rho = E[xy^*]$.

FUTURE TRENDS

As the above discussion indicates, there are numerous points where the current understanding is inadequate. The field is rich in opportunities for investigation of improved theory and techniques.

If one wants to improve on the methods described above, probably the best place to start will be to find ways to better incorporate *a priori* information into the procedure. A clear understanding of the problem and the nature of the data will often make the difference between a valuable and a useless analysis.

The use of higher-order cumulants as functions of higher-order moments which have the properties of correlations is increasing. Since cumulants above the second order are zero for Gaussian data, they may be a good way to filter out Gaussian noise in order to study non-Gaussian components. This use is handicapped by two problems. First, the probability distributions for the estimators are not as well understood. This makes testing of estimates, and estimation of false-alarm rates, difficult. This is aggravated by the fact that unless the sample size is very large, the random variability of the cumulant estimators is very large. Second, it is often not clear which cumulants to use. To date, the best innovations in this area seem to have consisted in clever identification of cumulants of interest.

The most useful data analysis techniques tend to be based on arguments from decision theory and/or game theory. Information theory has also played a role, primarily in the use of ideas about entropy. In the future, information theory will probably play a more important role. From this viewpoint, the binary decision problem, that is, the detection problem, seems well supported by convincing theoretical arguments. This is much less true for the multiple-hypothesis problem, that is the estimation problem. Occasionally, the basic ideas here should be carefully revisited.

ACKNOWLEDGMENTS

Much of the material on probability theory for complex variables was worked out on funds from the U.S. Office of Naval Research In-house Laboratory Independent Research program.

BIBLIOGRAPHY

1. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Function,* New York: Dover, 1972.
2. R. W. Hamming, *Digital Filters,* Englewood Cliffs, NJ: Prentice-Hall, Inc., 1977.
3. C. Eckart, *Optimal Rectifier Systems for the Detection of Steady Signals,* La Jolla, CA: University of California Marine Physical Laboratory of the Scripps Institution of Oceanography, 1952.
4. A. M. Mood and F. A. Graybill, *Introduction to the Theory of Statistics,* New York: McGraw-Hill, 1963.
5. N. R. Goodman, *On the joint estimation of the spectra, cospectrum, and quadrature spectrum of a two-dimensional stationary gaussian process,* Technical Report, Engineering Statistics Laboratory, New York University, 1957.

Reading List

A. Bertlesen, On non-null distributions connected with testing that a real normal distribution is complex, *J. Multivariate Anal.,* **32**: 282–289, 1990.

R. Fortet, *Elements of Probability Theory,* London: Gordon and Breach, 1977.

C. G. Khatri and C. D. Bhavsar, Some asymptotic inferential problems connected with complex elliptical distribution, *J. Multivariate Anal.,* **35**: 66–85, 1990.

C. L. Nikias and M. Shao, *Signal Processing with Alpha-Stable Distributions and Applications,* New York: Wiley, 1995.

B. Picinbono, On circularity, *IEEE Trans. Signal Process.,* **42**: 3473–3482, 1994.

A. K. Saxena, Complex multivariate statistical analysis: An annotated bibliography, *Int. Statist. Rev.,* **46**: 209–214, 1978.

R. A. Wooding, The multivariate distribution of comlpex normal variables, *Biometrika,* **43**: 212–215, 1956. Historical interest aside, this paper is interesting for the connection with Hilbert transforms.

DAVID J. EDELBLUTE
SPAWAR Systems Center San Diego