I. THE MAXIMUM ENTROPY VARIATIONAL PRINCIPLE FOR SINGLE CHANNEL POWER SPECTRAL ANALYSIS

In its most elementary and useful form, the maximum entropy variational principle for estimating the power spectrum of a single channel, stationary, complex time series can be stated as:

Find the power spectrum, $P(f)$, that maximizes the value of

$$\int_{-W}^{W} \ln P(f) \, df \qquad \text{(I-1)}$$

under the constraint that $P(f)$ satisfies a set of N linear functional measurement equations

$$\int_{-W}^{W} P(f) \, G_n(f) \, df = g_n \, , \quad n=1 \text{ to } N \, . \qquad \text{(I-2)}$$

The standard assumptions used in time series analysis are also assumed here. That is, the time series is sampled at a uniform period of Δt , $W = 1/(2\Delta t) =$ the Nyquist fold-over frequency and that the power spectrum of the time series is band-limited to $\pm W$. The $G_n(f)$ are the probe or test functions and the g_n are the resulting values of the measurements. It is also implicitly assumed that the constraint equations (I-2) are sufficient to place an upper bound on (I-1).

The basic assumption involved in maximum entropy spectral analysis is that the stationary time series being analyzed is the most random or the least predictable time series that is consistent with the measurements (I-2). In terms of information theory, this statement could be interpreted to mean that the entropy per sample of the time series is a maximum. For a given power spectrum, it is easily derived from theorems in Shannon and Weaver (1959)[3] that the maximum entropy time series is governed by a gaussian joint probability function and that the entropy is proportional to the integral of the logarithm of the spectrum. Thus, the maximum entropy stationary time series is the gaussian time series whose spectrum maximizes (I-1) under the constraint equations (I-2).

In retrospect, if one considers that we are simply estimating a spectrum from a few measurements of the second order statistics of a time series, one could ask why a probability distribution has entered into the picture. Actually, there is no need to make a gaussian assumption to relate (I-1) to the predictability of a time series. In fact, (I-1) is directly related to the least mean square error in predicting the next point of a time series by an infinitely long linear operator. Thus (I-1) has an importance that is completely independent of information theory and any gaussian assumptions. To stress this fact, we shall later on use (I-1) to derive the entropy of a gaussian time series.

The power spectrum of a stationary time series completely specifies the second order statistics of the random process. As an example of the information contained in the power spectrum, if we filter the time series with two arbitrary linear, time invariant digital filters with complex frequency responses $H_1(f)$ and $H_2(f)$ to get the two output time series $y_1(t)$ and $y_2(t)$, then the average value of $y_1^*(t)\, y_2(t)$ is given by

$$\overline{y_1^*(t)\; y_2(t)} \;=\; \int_{-W}^{W} P(f)\; H_1^*(f)\; H_2(f)\; df \;. \qquad \text{(I-3)}$$

Thus, the power spectrum contains all of the statistical information about the time series that is needed to find the average product of two linear, time invariant operations. This is of course the reason that power spectra are so important in the theory of linear, mean square error processing of stationary time series. On the other hand, without making any additional assumptions, the <u>only</u> thing the power spectrum tells us about the stationary time series is the value of integrals of the form (I -2) or (I -3). We should note that our linear functional constraint equations (I-2) belong to the class of measurements as expressed by (I-3).

Although knowledge of the second order statistics alone is insufficient for determining the entropy per sample of the time series, it does set an upper bound on the entropy per sample. This upper bound is achieved if the time series is a gaussian process. We shall now derive the entropy of a gaussian time series, remembering from Shannon and Weaver (1959)[3] that the entropy of a zero mean gaussian random variable with variance σ^2 is $\frac{1}{2} \ln(2\pi e\, \sigma^2)$.

Consider the situation in which we have received all of the samples up to x_0 and we wish to determine how much information we will receive when we find out the value of x_0 . If our time series consists of white gaussian noise so that each sample is statistically independent of every other sample, then the entropy of x_0 will be $\frac{1}{2} \ln(2\pi e\, P_0)$, where P_0 is the average square value of the zero mean time series. However, if the time series has a non-white spectrum, then x_0 will be at least partially predictable from the previous samples and its entropy will be less than $\frac{1}{2} \ln(2\pi e\, P_0)$.

It is known that for a gaussian process, the optimum predictor is a linear predictor. That is, the best prediction of x_0 from the previous samples is of the form $\sum_{n=-1}^{-\infty} a_n x_n$. In addition, the error in the prediction is gaussianly distributed with zero mean. Since the unpredictable part of x_0 contains the actual new information we will receive when the value of x_0 becomes known, the actual entropy of x_0 is $\frac{1}{2}\,\ell n(2\pi e\, P_\infty)$, where P_∞ is the least mean square error in predicting x_0 from the infinite set of previous samples. It is known[4] and will be proven later in chapter II-C that

$$\frac{1}{2W}\int_{-W}^{W} \ell n\; P(f)\; df \;=\; \ell n(P_\infty / 2W) \quad . \tag{I-4}$$

Thus, for a gaussian time series, the entropy per sample is given by

$$\frac{1}{4W}\int_{-W}^{W} \ell n(4\pi e\, W P(f))\; df \tag{I-5}$$

and this is the upper bound on the entropy of any time series whose power spectrum is $P(f)$.

While (I-5) depends on the gaussian assumption, (I-4) is valid for any stationary time series. Thus, <u>without</u> making any assumptions about probability distributions, one can say that any time series whose power spectrum maximizes (I-1) under the constraints (I-2) has the largest linear, least mean square prediction error. That is, the least mean square error in predicting x_0 by a linear combination of all previous samples is a maximum. In this sense, the time series will belong to the class of the most unpredictable or random time series that satisfy the constraint equations. If one now makes the assumption that the entropy of a time series is a non-decreasing function of its linear, least mean square prediction error, then maximizing (I-1) maximizes the entropy. It is easy to generate pairs of time series which do not obey this assumption. However, on an "everything else being equal" basis, the larger the mean square error, the larger the entropy. This assumption is much less stringent than the gaussian assumption. However, it is completely adequate for determining a spectrum which maximizes the entropy since it establishes a greater than or equal hierarchy among all spectra. Thus, if we have two spectra, $P_1(f)$ and $P_2(f)$, then if

$$\int_{-W}^{W} \ln P_1(f)\, df \geq \int_{-W}^{W} \ln P_2(f)\, df ,$$

then the entropy of $P_1(f)$ is greater than or equal to the entropy of $P_2(f)$. Restricting $P(f)$ to the set of spectra that satisfy (I-2), a maximum entropy spectrum can then be obtained.

Actually, since we are only attempting to estimate the power spectrum of the time series and our measurements (I-2) contain only information about the second order statistics, one should perhaps be satisfied with interpreting (I-1) as maximizing the linear, least mean square prediction error without extending the interpretation to entropy. An even more pragmatic viewpoint is to say that maximizing (I-1) is a straightforward variational problem with constraints. It seems to give good spectral estimates and that all of the above philosophical considerations do not change a single number in calculating the spectral estimate. Each point of view has its merits.