III-A. THE ENTROPY OF A MULTICHANNEL TIME SERIES

In order to derive the multichannel maximum entropy spectral analysis equations, we must first define the entropy per time step of a set of stationary time series. If our M time series are statistically independent and gaussianly distributed, then the entropy per sample of each channel is proportional to the integral of the logarithm of its spectrum and the entropy per time step for all M time series is the sum of the entropies. This assumes that the entropy per step of each channel is finite. See the end of this section for a discussion of the special case of multichannel entropies of minus infinity. If we ignore proportionality constants for the remainder of this section, then we can say that the multichannel entropy for independent channels is given by

$$\sum_{m=1}^{M} \int_{-W}^{W} \ln P_m(f)\, df \;=\; \int_{-W}^{W} \ln \prod_{m=1}^{M} P_m(f)\, df \;=\; \int_{-W}^{W} \ln\,[\det P(f)]\, df \;, \qquad \text{(III-1)}$$

where $P(f)$ is the multichannel power spectrum matrix. In this case, $P(f)$ is M by M and diagonal with its diagonal elements being given by $P_m(f)$, $m=1$ to M . We will now show that the integral of the logarithm of the determinant of $P(f)$ is a reasonable definition for the entropy per time step of a multichannel spectrum.

Suppose we have a set of M gaussianly distributed random variables, v_m , with a general positive definite covariance matrix, R . If the variables were independent, i.e., if R were diagonal, then the entropy of the set of variables would be the sum of the logarithms of their variances, or equivalently, the logarithm of the determinant of R . However, if the random variables are not independent, we can change them to a new set which are independent by using prediction error filters. For example, we can make our new set of variables be: (1) v_M ; (2) the error in predicting v_{M-1} from v_M ; (3) the error in predicting v_{M-2} from v_{M-1} and v_M ; etc., until we finally get for our last new variable the error in predicting v_1 from all the other variables. This set of prediction error variables will be independent and their entropy has been defined. It is seen that this procedure is analogous to that used to convert the correlated samples of a colored time series to the independent samples of a whitened time series. In matrix form, our "whitening" of the M by M covariance matrix looks like

$$U^T R U = D ,$$

where D is diagonal with positive elements d_m and U is the solution of the equation below (shown explicitly for M=4).

$$RU \equiv R \begin{bmatrix} 1 & 0 & 0 & 0 \\ * & 1 & 0 & 0 \\ * & * & 1 & 0 \\ * & * & * & 1 \end{bmatrix} = \begin{bmatrix} d_1 & * & * & * \\ 0 & d_2 & * & * \\ 0 & 0 & d_3 & * \\ 0 & 0 & 0 & d_4 \end{bmatrix} , \qquad \text{(III-2)}$$

where * indicates an element to be solved for. The first column of U is the filter for obtaining the least mean square error in predicting v_1 from v_m , m=2 to M , etc. We note that $|U| = 1$, so that $|D| = |R|$. Thus, the logarithm of the determinant of the covariance matrix is invariant under the U transformation. This should seem reasonable since the unity weight on the predicted variable of the prediction error filter does not gain up the variable. Thus, since the U matrix is reversible and does not scale the variables, one can believe that it should not change the entropy. Using this result, if the multichannel power spectrum matrix is not diagonal but is constant with frequency, its entropy will be given by (III-1). However, we want to show that (III-1) has broad applicability and to do this we call on multichannel prediction error filters. An N long multichannel prediction error filter has a matrix z transform of

$$F(z) = I + A_1 z + A_2 z^2 + \ldots + A_N z^N ,$$

where the leading matrix is the identity matrix and A_n , $n=1$ to N , are M by M matrices. The multichannel prediction error filter actually consists of M filters, where the mth filter predicts the next point on the mth channel from the past multichannel data. Because of the unity weight on the predicted point, the variables are not scaled and the entropy should not be changed.

Just like the single channel prediction error filter, the multichannel prediction error filter can "whiten" a multichannel time series, that is, convert the power spectral matrix to a constant matrix (but not necessarily a diagonal matrix). If this constant power spectral matrix is positive definite, then the multichannel prediction error filter is minimum phase and we find that the z polynomial given by the determinant of $F(z)$ has all of its roots outside the unit circle. Furthermore, the coefficient of z^0 is unity. Thus, using the single channel theorem, we have that

$$\int_{-W}^{+W} \ln |F(z)| \, df = 0 ,$$

where we have written $|F(z)|$ for the determinant of $F(z)$. Now , if we have a multichannel time series with a constant positive definite spectral matrix P and filter it with $F^{-1}(z)$, we obtain a new spectral matrix given by $F^{-1\dagger}(z^{-1}) P F^{-1}(z)$, where the dagger indicated the complex conjugate transpose operation. However, we note that

$$\int_{-W}^{W} \ln \left|F^{-1\dagger}(z^{-1})\, P\, F^{-1}(z)\right| df = -\int_{-W}^{W} \ln \left|F^{\dagger}(z^{-1})\right| \, df + \int_{-W}^{W} \ln|P| \, df$$

$$- \int_{-W}^{W} \ln \left|F(z)\right| \, df = \int_{-W}^{W} \ln |P| \, df .$$

Thus, the multichannel spectrum $F^{-1\dagger}(z^{-1})\, P\, F^{-1}(z)$ has the same entropy as P . If (III-1) is finite for a multichannel power spectral matrix $P(f)$, then $P(f)$ can be expressed as $F^{-1\dagger}(z^{-1})\ P\ F^{-1}(z)$ where $F(z)$ is a probably infinitely long multichannel prediction error filter and P is a positive definite constant spectral matrix whose entropy is the same as $P(f)$. Thus, when (III-1) is finite, it is a reasonable definition for the entropy of a multichannel time series.

If (III-1) is minus infinity, then this means that at least one of the channels can be predicted perfectly from present and past multichannel values. If a channel has such a deterministic relationship to the other channels, then it should be removed from the spectral matrix since its actual information content is zero. This is clearly a logical flaw in the use of (III-1) as the definition of entropy. However, since we are mainly concerned here with developing a variational principle, we shall not go deeper into this problem.