AN INFORMATION THEORETIC PROOF OF BURG'S
MAXIMUM ENTROPY SPECTRUM

by
B. S. Choi
and
Thomas Cover

TECHNICAL REPORT NO. 49
OCTOBER 7, 1983

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT
ECS82-11568

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
AN INFORMATION THEORETIC PROOF OF
BURG'S MAXIMUM ENTROPY SPECTRUM

by

B.S. Choi
Thomas M. Cover

TECHNICAL REPORT NO. 49
OCTOBER 7, 1983

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION
GRANT ECS82-11568

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
An Information Theoretic Proof of
Burg's Maximum Entropy Spectrum

by

B.S. Choi†
Thomas M. Cover††

ABSTRACT

There are now many proofs that the maximum entropy stationary stochastic process, subject to a finite number of covariance constraints, is the Gauss Markov process of appropriate order. The associated spectrum is Burg's maximum entropy spectral density. We pose a somewhat broader entropy maximization problem, in which stationarity, for example, is not assumed, and shift the burden of proof from the previous focus on the calculus of variations and time series techniques to a string of information theoretic inequalities. This results in an elementary proof.

† Stanford University, Department of Statistics.
†† Stanford University, Departments of Statistics and Electrical Engineering. This work was partially supported by National Science Foundation Grant ECS83-11569 and Joint Services Electronics Program DAAG29-81-K-0057.
1. Preliminaries.

We shall give some necessary definitions and go directly to a proof of the characterization of the maximum entropy stochastic process given covariance constraints. Section 3 has the history.

Let \( \{X_i\}_{i=1}^\infty \) be a stochastic process specified by its marginal probability density functions \( f(x_1, x_2, \ldots, x_n) \), \( n = 1, 2, \ldots \). Then the (differential) entropy of the n-sequence \( X_1, X_2, \ldots, X_n \) is defined by

\[
h(X_1, X_2, \ldots, X_n) = - \int f(x_1, \ldots, x_n) \ln f(x_1, \ldots, x_n) \, dx_1 \, dx_2 \cdots \, dx_n.
\]  

\( = h(f) \).

The stochastic process \( \{X_i\} \) will be said to have an entropy rate

\[
h = \lim_{n \to \infty} \frac{h(X_1, X_2, \ldots, X_n)}{n}
\]  

if the limit exists. It is known that the limit always exists for stationary processes.
2. The Proof.

We prove the following theorem:

**Theorem 1:** The stochastic process \( \{X_i\}_{i=1}^{\infty} \) that maximizes the differential entropy rate \( h \) subject to the autocorrelation constraints

\[
E X_i X_{i+k} = \alpha_k, \quad k = 0,1,2,\ldots,p, \quad i = 1,2,\ldots
\]

is the minimal order Gauss Markov process satisfying these constraints.

**Remark:** This \( p \)-th order Gauss Markov process simultaneously solves the maximization problems

\[
\max_{n} \frac{h(X_1,X_2,\ldots,X_n)}{n}, \quad n = 1,2,\ldots,
\]

subject to the above autocorrelation constraints.

**Proof of Theorem:** Let \( X_1,X_2,\ldots,X_n \) be any collection of random variables satisfying (3). Let \( Z_1,Z_2,\ldots,Z_n \) be zero mean multivariate normal with the same covariance as \( X_1,X_2,\ldots,X_n \). And let \( Z'_1,Z'_2,\ldots,Z'_n \) be the \( p \)-th order Gauss Markov process satisfying the covariance constraints in (4). Then, for \( n \geq p \),

\[
h(X_1,\ldots,X_n) \overset{(a)}{\leq} h(Z_1,Z_2,\ldots,Z_n)
\]

\[
\overset{(b)}{=} h(Z_1,Z_2,\ldots,Z_p) + \sum_{k=p+1}^{n} h(Z_k | Z_{k-1},\ldots,Z_1)
\]

\[
\overset{(c)}{=} h(Z_1,Z_2,\ldots,Z_p) + \sum_{k=p+1}^{n} h(Z_k | Z_{k-1},Z_{k-2},\ldots,Z_{k-p})
\]

\[
= h(Z'_1,Z'_2,\ldots,Z'_n)
\]

(4)

Here (b) is the chain rule for entropy, and inequality (c) follows from
\( h(A \mid B, C) \leq h(A \mid B) \). (See standard texts like Ash [1] and Gallager [2].) Inequality (a) follows from the information inequality. See Section 3. Thus the \( p \)-th order Gauss Markov process \( Z'_1, Z'_2, \ldots, Z'_n \) with covariances \( \alpha_0, \alpha_1, \ldots, \alpha_p \) has higher entropy \( h(Z'_1, Z'_2, \ldots, Z'_n) \) than any other process satisfying the autocorrelation constraints. Consequently,

\[
\lim_{n \to \infty} \frac{1}{n} h(X_1, \ldots, X_n) \leq \frac{1}{n} h(Z'_1, \ldots, Z'_n) = h, \tag{5}
\]

for all stochastic processes \( \{X_i\} \) satisfying the covariance constraints, thus proving the theorem.
3. **Comments on Proof.**

For completeness, we shall provide a proof of the well known inequality (a) in the proof of the theorem. See for example, Berger [3]. Let \( f(x_1, \ldots, x_n) \) be a probability density, and let

\[
\phi(x) = \frac{1}{(2\pi)^{n/2} \ |K|^{1/2}} e^{-x^t K^{-1} x} \tag{6}
\]

be the \( n \)-variate normal density with covariance matrix

\[
K = \int x x^t f(x) dx .
\]

Let \( D(f \| g) = \int \int \ln \frac{f}{g} \) denote the Kullback-Leibler information number for \( f \) relative to \( g \). It is known from Jensen's inequality that \( D(f \| g) \geq 0 \). Thus

\[
0 \leq D(f \| \phi) \triangleq \int \int f(x) \ln \frac{f(x)}{\phi(x)} dx \\
= \int \int \ln f - \int \int \ln \phi 
\tag{7}
\]

But

\[
\int \int \ln \phi = \int \phi \ln \phi \tag{8}
\]

because both are expectations of quadratic forms in \( x \). These expected quadratic forms are completely determined by the covariances in (3), and are thus equal. Substituting (8) in (7), we have

\[
0 \leq -h(f) - \int \int \ln \phi \\
= -h(f) - \int \phi \ln \phi
\]

and
\[ h(f) \leq h(\phi), \quad (9) \]
as desired. This completes the proof of inequality (c).

Remark: A pleasing byproduct of the proof is that the solutions to all of the finite-dimensional maximization problems, and therefore, of the (limiting) entropy rate maximization problem, are given by the finite dimensional marginal densities \( f(x_1, x_2, \ldots, x_n), \ n = 1, 2, \ldots \), of a single stochastic process — the Gauss Markov process of order \( p \).
4. Equivalent Characterizations of the Solution.

Now that the maximum entropy process has been characterized it is simple to provide an equivalent characterization.

We shall give the autoregressive characterization of the maximizing process via the Yule-Walker equations. Let

\[ X_n = - \sum_{i=1}^{p} a_i X_{n-i} + Z_n , \]  

(10)

where \( Z_1, Z_2, \ldots \) are i.i.d. \( \sim N(0, \sigma^2) \), and \( a_1, a_2, \ldots, a_p \) and \( \sigma^2 \) are to be determined. The autocorrelation constraints (3) determine the coefficients \( a_1, a_2, \ldots, a_p \) according to the Yule-Walker equations

\[ \sum_{j=0}^{p} a_j \alpha_{l-j} = 0 , \quad l = 1, 2, \ldots, p \]

(11)

where \( a_0 = 1 \), and \( \sigma^2 \) is given by

\[ \sigma^2 = \sum_{j=0}^{p} a_j \alpha_j . \]

(12)

Let \( (X_1, X_2, \ldots, X_p) \) be zero mean multivariate normal with covariance matrix given by (3). Then let \( X_n \), \( n \geq p \), be determined by the autoregressive equation ( ). Inspection of ( ) yields the remaining autocovariance values

\[ \alpha_l = \sum_{j=1}^{p} a_j \alpha_{l-j} , \quad l \geq p + 1 . \]

(13)

Thus, as was observed by Burg, the maximum entropy stochastic process is not obtained by setting the unspecified covariance terms equal to zero, but instead is given by letting the \( p \)-th order autoregressive process "run" according to the Yule-Walker equations.

Finally, taking the Fourier transform of \( \alpha_0, \alpha_1, \ldots \) given in (3) and ( ) yields
$$S(\lambda) = \frac{1}{2\pi} \sum_{l=-\infty}^{\infty} \alpha_l e^{-i\lambda l}$$

$$= \frac{\sigma^2}{2\pi} \frac{1}{| \sum_{j=0}^{p} a_j e^{i\lambda j} |^2}.$$  \hspace{1cm} (14)

This is Burg's maximum entropy spectral density subject to the covariance constraints $\alpha_0, \alpha_1, \ldots, \alpha_p$.

Finally, the resulting maximum entropy rate is

$$h = \lim_{n \to \infty} \frac{1}{n} \left[ h(X_1, \ldots, x_p) + \sum_{l=p+1}^{n} h(X_l | X_{l-1}, \ldots, X_{l-p}) \right]$$

$$= h(X_{p+1} | X_p, X_{p-1}, \ldots, X_1)$$

$$= \frac{1}{2} \ln \left( 2\pi e\sigma^2 \right),$$ \hspace{1cm} (15)

where $\sigma^2$ is given in (2).
5. **History.**

Burg (1967) introduced the maximum entropy spectral density by exhibiting the solution to the problem of maximizing the entropy rate

\[
h = \frac{1}{2} \ln(2\pi e) + \frac{1}{4\pi} \int_{-\pi}^{\pi} \ln(2\pi S(\lambda)) d\lambda
\]

(16)

of a Gaussian stochastic process

\[
S(\lambda) = \frac{1}{2\pi} \sum_{l=-\infty}^{\infty} \sigma(l) e^{-i\lambda l},
\]

(17)

and \( \{\sigma(l)\}_{l=-\infty}^{\infty} \) is an arbitrary autocovariance function subject to the constraints

\[
\sigma(0) = \alpha_0, \sigma(1) = \alpha_1, \ldots, \sigma(p) = \alpha_p.
\]

(18)

Proof that the autoregressive process spectral density is the maximum entropy spectral density has been established by variational methods by Smylie, Clarke and Ulrych (1973, pp. 402-419), using the Lagrange multiplier method, and independently by Edward and Fitelson (1973). Burg (1975), Ulrych and Bishop (1975), Haykin and Kesler (1979, pp. 16-21), and Robinson (1982) follow Smylie's method; and Ulrych and Ooe (1979) and McDonough (1979) use Edward's.

The calculus of variations necessary to show that

\[
S^*(\lambda) = \frac{\sigma^2}{2\pi} \frac{1}{\left| \sum_{j=0}^{p} a_j e^{i\lambda j} \right|^2}
\]

(19)

is the solution to (16) is tricky. Indeed Burg (1975) and Smylie (1973) show only that the first variation about \( S^*(\lambda) \) is zero.

Van den Bos (1971) maximizes the entropy \( h(x_1, x_2, \ldots, x_{p+1}) \) subject to the constraints \( \sigma(0) = \alpha_0, \sigma(1) = \alpha_1, \ldots, \sigma(p) = \alpha_p \) by differential calculus,
but further argument is required to extend his solution to the maximization of
\( h(X_1,\ldots,X_n) \), \( n > p + 1 \).

Akaike (1977), maximizes another form of the entropy rate \( h \), i.e.,
\[
    h = \frac{1}{2} \log (2\pi e) + \frac{1}{2} \text{Var} (\epsilon_t), \tag{20}
\]
where \( \epsilon_t \) is the prediction error of the best linear predictor of \( x_t \) in terms of all
the past \( x_{t-1}, x_{t-2}, \ldots \). Equation (20) can be derived from (16) through
Kolmogorov's equality (1941):
\[
    \text{Var}(\epsilon_t) = 2\pi \exp \left[ \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln S_x(\lambda) d\lambda \right].
\]
Using prediction theory, it is shown that \( \text{Var}(\epsilon_t) \) has its maximum if
\[
    \epsilon_t = x_t + a_1 x_{t-1} + a_2 x_{t-2} + \cdots + a_p x_{t-p}.
\]
where \( a_1, a_2, \ldots, a_p \) are given in ( ).

With hindsight, we see that all of the maximization can be captured in the
information theoretic string of inequalities in Equation (5) of Theorem 1, and
that the global maximality of \( S^*(\lambda) \) follows automatically from verifying that
\( S^*(\lambda) \) is the spectrum of the process specified by the theorem.
6. Conclusions.

A bare bones summary of the proof is that the entropy of a finite segment of a stochastic process is bounded above by the entropy of a segment of a Gaussian random process with the same covariance structure. This entropy is in turn bounded above by the entropy of the minimal order Gauss Markov process satisfying the given covariance constraints. Such a process exists and has a convenient characterization via the Yule-Walker equations. Thus the maximum entropy stochastic process is obtained.

We wish to mention that the maximum entropy spectrum actually arises as the answer to a certain "physical" question. Suppose $X_1, X_2, \ldots$ are independent identically distributed uniform random variables. Suppose also that the following empirical covariance constraints are observed:

$$\frac{1}{n} \sum_{i=1}^{n} X_i X_{i+k} = \alpha_k , \quad k = 0, 1, \ldots, p .$$

(21)

What is the conditional distribution on $(X_1, X_2, \ldots, X_m)$? It will be shown in Choi and Cover (1983) that the limit, as $n \to \infty$, of the conditional densities given the empirical constraint (21) tends to the maximum entropy process specified in Theorem 1. Thus, an independent uniform process conditioned on empirical correlations looks like a Gauss Markov process.
References


- 14 -