A PROOF OF BURG'S THEOREM*

B. S. Choi

Department of Applied Statistics, Yonsei University, Seoul, Korea

Thomas M. Cover

Departments of Statistics and Electrical Engineering, Stanford University, Stanford, CA 94305

There are now many proofs that the maximum entropy stationary stochastic process, subject to a finite number of autocorrelation 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 a simple proof.

---

*Expanded version of a paper published originally in Proceedings of the IEEE 72, pp. 1094-1095 (1984).

C. R. Smith and G. J. Erickson (eds.),
Maximum-Entropy and Bayesian Spectral Analysis and Estimation Problems, 75–84

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 5 has the history. In the concluding section, we mention a conditional limiting characterization of Gauss Markov processes.

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

$$h(X_1,X_2,\ldots,X_n) = -\int f(x_1,\ldots,x_n) \ln f(x_1,\ldots,x_n)\, dx_1\, dx_2 \ldots dx_n$$

$$= h(f) . \tag{1}$$

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

$$h = \lim_{n\to\infty} \frac{h(X_1,X_2,\ldots,X_n)}{n} \tag{2}$$

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

$$E\, X_i\, X_{i+k} = \alpha_k , \quad k = 0,1,2,\ldots,p, \quad i = 1,2,\ldots, \tag{3}$$

is the minimal order Gauss Markov process satisfying these constraints.

Remark: This pth order Gauss Markov process simultaneously solves the maximization problems

$$\max \frac{h(X_1,X_2,\ldots,X_n)}{n} , \quad n = 1,2,\ldots, \tag{4}$$

subject to the above autocorrelation constraints.

Proof: Let $X_1,X_2,\ldots,X_n$ be any collection of random variables satisfying Eq. (3). Let $Z_1,Z_2,\ldots,Z_n$ be zero mean multivariate normal with a covariance matrix given by the correlation matrix of $X_1,X_2,\ldots,X_n$. And let $Z_1',Z_2',\ldots,Z_n'$ be the pth order Gauss Markov process with covariance specified in Eq. (3). Then, for $n \geq p$,

$$h(X_1,\ldots,X_n) \leq h(Z_1,Z_2,\ldots,Z_n) \tag{5a}$$

$$= h(Z_1,Z_2,\ldots,Z_p) + \sum_{k=p+1}^{n} h(Z_k \mid Z_{k-1},\ldots,Z_1) \tag{5b}$$

$$\leq h(Z_1,Z_2,\ldots,Z_p) + \sum_{k=p+1}^{n} h(Z_k \mid Z_{k-1},Z_{k-2},\ldots,Z_{k-p}) \tag{5c}$$

$$= h(Z_1',Z_2',\ldots,Z_p') + \sum_{k=p+1}^{n} h(Z_k' \mid Z_{k-1}',\ldots,Z_{k-p}') \tag{5d}$$

$$= h(Z_1',Z_2',\ldots,Z_n') \,. \tag{5e}$$

Here inequality (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, 1965, and Gallager, 1968.] Inequality (a) follows from the information inequality, as shown in Section 3. Thus the pth 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 $\alpha_0,\alpha_1,\ldots,\alpha_p$. Consequently,

$$\lim_{n\to\infty} \frac{1}{n} h(X_1,\ldots,X_n) \leq \lim_{n\to\infty} \frac{1}{n} h(Z_1',\ldots,Z_n') = h \,, \tag{6}$$

for all stochastic processes $\{X_i\}$ satisfying the covariance constraints, thus proving the theorem.

3. Comments on the Proof

For completeness, we provide a proof of the well known inequality (a) in the proof of Theorem 1. [See for example Berger, 1971.] Let $f(x_1,\ldots,x_n)$ be a probability density function, and let

$$\phi(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\,|K|^{1/2}}\, e^{-x^t K^{-1} x/2} \tag{7}$$

be the n-variate normal probability density with covariance matrix

$$K = \int \mathbf{x}\,\mathbf{x}^t\, f(\mathbf{x})\, d\mathbf{x} \,. \tag{8}$$

Thus, ϕ and f have the same correlation matrix K.

Let

$$D(f\|g) = \int f \ln \frac{f}{g} \tag{9}$$

denote the Kullback-Leibler information number for f relative to g. It is known from Jensen's inequality that $D(f\|g) \geq 0$ for any probability densities f and g. Thus,

$$0 \leq D(f\|\phi) \stackrel{\Delta}{=} \int f(\mathbf{x}) \ln \frac{f(\mathbf{x})}{\phi(\mathbf{x})} d\mathbf{x}$$

$$= \int f \ln f - \int f \ln \phi \, . \tag{10}$$

But

$$\int f \ln \phi = \int \phi \ln \phi \tag{11}$$

because both are expectations of quadratic forms in $\mathbf{x}$. These expected quadratic forms are completely determined by Eq. (3), and are thus equal. Substituting Eq. (11) into Eq. (10) and using Eq. (1), we have

$$0 \leq -h(f) - \int f \ln \phi$$

$$= -h(f) - \int \phi \ln \phi$$

$$= -h(f) + h(\phi) \tag{12}$$

and

$$h(f) \leq h(\phi) \, , \tag{13}$$

as desired. This completes the proof of inequality (5a).

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 Solutions

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 maximum entropy process by means of the Yule-Walker equations. If the $p \times p$ symmetric Toeplitz matrix whose (i,j)th element is $\alpha_{|i-j|}$ is positive definite, then there exists a unique solution set $\{a_1,\ldots,a_p\}$ of the Yule-Walker equations

$$\sum_{i=0}^{p} a_i \alpha_{|\ell-i|} = 0, \qquad \ell = 1,\ldots,p, \tag{14}$$

where $a_0 = 1$. And then it can be proved [Choi, 1983] that $\Sigma_{i=0}^{p} a_i \alpha_i$ is positive. Thus, we can define

$$\sigma^2 = \sum_{i=0}^{p} a_i \alpha_i . \tag{15}$$

Consider the corresponding autoregressive process $\{X_n\}$ of order p,

$$X_n = -\sum_{i=1}^{p} a_i X_{n-i} + Z_n , \tag{16}$$

where $Z_1, Z_2, \ldots$ are independent and identically distributed normal random variables with mean 0 and variance σ^2. Inspection of Eqs. (3) and (14) yields the remaining autocovariance values

$$\alpha_\ell = \sum_{j=1}^{p} a_j \alpha_{|\ell-j|} , \qquad \ell \geq p+1 . \tag{17}$$

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 pth order autoregressive process "run" according to the Yule-Walker equations.

Finally, taking the Fourier transform of $\alpha_0, \alpha_1, \ldots$ given in Eqs. (3) and (17), yields the spectral density $S(\lambda)$:

$$S(\lambda) = \frac{1}{2\pi} \sum_{\ell=-\infty}^{\infty} \alpha_\ell e^{-i\lambda\ell}$$

$$= \frac{\sigma^2}{2\pi} \frac{1}{\left| \sum_{j=0}^{p} a_j e^{i\lambda j} \right|^2} . \tag{18}$$

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

The resulting maximum entropy rate is

$$h = \lim_{n\to\infty} \frac{1}{n} \left[h(X_1, \ldots, X_p) + \sum_{\ell=p+1}^{n} h(X_\ell \mid X_{\ell-1}, \ldots, X_{\ell-p}) \right]$$

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

$$= \frac{1}{2} \ln (2\pi e \sigma^2) , \tag{19}$$

where σ^2 is given in Eq. (15). Incidentally, the maximum entropy process will be less than pth order, although still determined by Eqs. (14), (15), (16), if Eq. (3) is not strictly positive definite. The true order of the process is the largest k for which $[\alpha_{|i-j|}]_{1 \le i,j \le k}$ is positive definite.

5. History

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

$$h = \frac{1}{2} \ln(2\pi e) + \frac{1}{4\pi} \int_{-\pi}^{\pi} \ln [2\pi S(\lambda)] \, d\lambda \tag{20}$$

where

$$S(\lambda) = \frac{1}{2\pi} \sum_{\ell=-\infty}^{\infty} \sigma(\ell) e^{-i\lambda\ell} , \tag{21}$$

and $\{\sigma(\ell)\}_{\ell=-\infty}^{\infty}$ is an arbitrary autocovariance function satisfying the constraints

$$\sigma(0) = \alpha_0 , \quad \sigma(1) = \alpha_1 , \quad \ldots , \quad \sigma(p) = \alpha_p . \tag{22}$$

Proof that the pth order Gaussian 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. Ulrych and Ooe [1979] and McDonough [1979] use Edward's method. See also Grandell et al. [1980].

The calculus of variations necessary to show that

$$S^*(\lambda) = \frac{\sigma^2}{2\pi} \frac{1}{\left| \sum_{j=0}^{p} a_j e^{i\lambda j} \right|^2} \tag{23}$$

is the solution to Eq. (20) is tricky. Smylie et al. [1973] show that the first variation about $S^*(\lambda)$ is zero. Further considerations establish S^* as a maximum.

Van den Bos [1971] maximizes the entropy $h(x_1,x_2,\ldots,x_{p+1})$ subject to the constraints (22) by differential calculus, but further argument is required to extend his solution to the maximization of $h(X_1,\ldots,X_n)$, $n > p+1$. Feder and Weinstein [1984] have carried this out.

Akaike [1977] maximizes another form of the entropy rate h, that is,

$$h = \tfrac{1}{2} \log(2\pi e) + \tfrac{1}{2} \operatorname{Var}(\varepsilon_t), \tag{24}$$

where ε_t is the prediction error of the best linear predictor of X_t in terms of all the past $X_{t-1},X_{t-2},\ldots$. Of course, Eq. (24) holds only if the process is Gaussian. Equation (24) can be derived from Eq. (20) through Kolmogorov's equality [1941]:

$$\operatorname{Var}(\varepsilon_t) = 2\pi \exp\left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \ln S_X(\lambda)\, d\lambda \right]. \tag{25}$$

Using prediction theory, one can show that $\operatorname{Var}(\varepsilon_t)$ has its maximum if

$$\varepsilon_t = X_t + a_1 X_{t-1} + a_2 X_{t-2} + \cdots + a_p X_{t-p}, \tag{26}$$

where $a_1,a_2,\ldots,a_p$ are given in Eq. (14). For details, see Priestley [1981, pp. 604-606].

More details of proofs in this section can be found in Choi [1983].

With hindsight, we see that all of the maximization can be captured in the information theoretic string of inequalities in Eq. (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 by means of the Yule-Walker equations. Thus the maximum entropy stochastic process is obtained.

We 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:

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_{i+k} = \alpha_k, \quad k = 0,1,\ldots,p. \tag{27}$$

What is the conditional distribution on $(X_1, X_2, \ldots, X_m)$? It is shown in Choi and Cover [1987] that the limit, as $n \to \infty$, of the conditional probability densities given the empirical constraint (27) tends to the unconditional probability density function of the maximum entropy process specified in Theorem 1. Thus, an independent uniform process conditioned on empirical correlations looks like a Gauss Markov process.

7. Acknowledgments

This work was partially supported by National Science Foundation Grant ECS82-11568 and Joint Services Electronics Program DAAG29-81-K-0057.

A shortened version of this paper appears as a Letter in the Proceedings of the IEEE [Choi and Cover, 1984].

8. References

Akaike, H. (1977), "An entropy maximization principle," in P. Krishnaiah, ed., Proceedings of the Symposium on Applied Statistics, North-Holland, Amsterdam.

Ash, R. (1965), Information Theory, Wiley Interscience, New York.

Berger, T. (1971), Rate Distortion Theory, A Mathematical Basis for Data Compression, Prentice-Hall, N.J.

Burg, J. P. (1967), "Maximum entropy spectral analysis," presented at the 37th Meeting of the Society of Exploration Geophysicists; reprinted in D. G. Childers, ed. (1978), Modern Spectrum Analysis, IEEE Press, pp. 34-41.

Burg, J. P. (1975), "Maximum Entropy Spectral Analysis," Ph.D. dissertation, Department of Geophysics, Stanford University, Stanford, Calif.

Choi, B. S. (1983), "A Conditional Limit Characterization of the Maximum Entropy Spectral Density in Time Series Analysis," Ph.D. dissertation, Statistics Department, Stanford University.

Choi, B. S., and T. M. Cover (1984), "An information-theoretic proof of Burg's maximum entropy spectrum" (letter), Proc. IEEE 72, pp. 1094-1095.

Choi, B. S., and T. M. Cover (1987), "A conditional limit characterization of Gauss Markov processes," submitted to JASA.

Edward, J. A., and M. M. Fitelson (1973), "Notes on maximum-entropy processing," IEEE Trans. Inf. Theory IT-19, pp. 232-234; reprinted in D. G. Childers, ed. (1978), Modern Spectrum Analysis, IEEE Press, pp. 94-96.

Feder, M., and E. Weinstein (1984), "On the finite maximum entropy extrapolation," Proc. IEEE 72, pp. 1660-1662.

Gallager, R. (1968), Information Theory and Reliable Communication, Wiley, New York.

Grandell, J., M. Hamrud, and P. Toll (1980), "A remark on the correspondence between the maximum entropy method and the autoregressive model," IEEE Trans. Inf. Theory IT-26, pp. 750-751.

Haykin, S., and S. Kesler (1979), "Prediction-error filtering and maximum entropy spectral estimation," in S. Haykin, ed., Nonlinear Methods of Spectral Analysis, Springer, New York, pp. 9-72.

Kolmogorov, A. N. (1941), "Interpolation und Extrapolation von Stationären Zufälligen Folgen," Bull. Acad. Sci. URSS, Ser. Math. 5, pp. 3-41.

McDonough, R. N. (1979), "Application of the maximum-likelihood method and the maximum entropy method to array processing," in S. Haykin, ed., Nonlinear Methods of Spectral Analysis, Springer, New York, pp. 181-244.

Priestley, M. B. (1981), Spectral Analysis and Time Series, Vol. 1, Academic Press, New York.

Robinson, E. A. (1982), "A historical perspective of spectrum estimation," Proc. IEEE 70, pp. 885-907.

Smylie, D. G., G. K. C. Clarke, and T. J. Ulrych (1973), "Analysis of irregularities in the earth's rotation," Meth. Comp. Phys. 13, pp. 391-430.

Ulrych, T., and T. Bishop (1975), "Maximum entropy spectral analysis and autoregressive decomposition," Rev. Geophys. and Space Phys., 13, pp. 183-200; reprinted in D. G. Childers, ed. (1978), Modern Spectrum Analysis, IEEE Press, pp. 54-71.

Ulrych, T., and M. Ooe (1979), "Autoregressive and mixed autoregressive-moving average models and spectra," in S. Haykin, ed., Nonlinear Methods of Spectral Analysis, Springer, New York, pp. 73-126.

Van den Bos, A. (1971), "Alternative interpretation of maximum entropy spectral analysis," IEEE Trans. Inf. Theory IT-17, pp. 493-494; reprinted in D. G. Childers, ed. (1978), Modern Spectrum Analysis, IEEE Press, pp. 92-93.