MIMEOGRAPH SERIES #1632

APPLICATION OF MAXIMUM ENTROPY AND MINIMUM CROSS-ENTROPY FORMALISMS TO STOCHASTIC MODELING OF COMPLEX DYNAMIC SYSTEMS; FORMULATION OF THE PROBLEM.

Harvey J. Gold

Biomathematics Series No. 14
Institute of Statistics Mimeo Series No. 1632
North Carolina State University
Raleigh, North Carolina

September 1983

APPLICATION OF MAXIMUM ENTROPY AND MINIMUM CROSS-ENTROPY FORMALISMS TO STOCHASTIC MODELING OF COMPLEX DYNAMIC SYSTEMS FORMULATION OF THE PROBLEM.

SUMMARY

The general problem addressed in this paper is that of using data and observations to infer a probability distribution on the state space of a system and of making inferences concerning the time dynamics of that probability distribution. The goal is to make full use of the available information, but without imputation of "information" which we do not in fact have. The approach is through extension of the information theoretic methods of maximum entropy and minimum cross-entropy. These methods have shown themselves to be powerful tools applied to inferences on stationary probability distributions. The proposed approach forextension to dynamic systems rests on introduction of additional relationships drawn from theory of dynamic system modeling. The methodology is intended as a tool in the stochastic modeling of biological systems, with special emphasis on application to ecology and to environmental management.

1. INTRODUCTION

The purpose of this paper is to suggest an extension of the formalism of maximum entropy and minimum cross entropy methodology so as to be applicable to stochastic modeling of biological systems. For this it is necessary that the methodology be extended to a dynamic framework, and that proper account be taken of sources of randomness that have been neglected or given only limited attention in the literature.

The general problem is that of using data and observations to infer a probability distribution on the state space of a system, and to make inferences concerning the time dynamics of that probability distribution. The methodology is intended for use with stochastic models in population biology, with application to ecology and environmental management. The approach builds upon the maximum entropy principle of Jaynes (see, for example, Jaynes 1968), the information divergence measure of Kullback (1959; see also Shore and Johnson 1980), and the concepts of system morphism as developed by Zeigler (1976).

Section 2 contains a formal statement of the basic problem in the simplest stationary case and briefly reviews its solution in terms of the maximum entropy formalism. This solution is seen to rest upon important assumptions that weaken its relevance to biological applications. These are discussed in Subsections 2.1, 2.2, and 2.3. Section 3 discusses an approach for extending the formalism to include treatment of dynamic systems; that is, systems characterized by non-stationary probability distributions and observations at multiple times.

2. FORMULATION OF THE BASIC PROBLEM.

We begin by assuming two measure spaces U and M which represent two different ways of defining the state space of a given system. We designate them as the macrostate space (M-space) and the microstate space (U-space). The spaces are assumed to be related to each other in the general sense of strata of an abstraction hierarchy (Mesarovic, Macko and Takahara 1970). We will often take the M-space to refer to a space of observables, and the U-space to refer to a space of underlying, possibly non-observable variables. In the simplest cases, the macrostates may be considered to be aggregates of the microstates. In another common use, the macrospace is a space of summary statistics which describe salient characteristics of a set of numbers.

Some examples are:

a) M-states defined by mean and variance of a set of numbers, say the results of rolls of a die; U-states defined by the complete set of results (example discussed by Jaynes, 1978).

b) M-states defined by thermodynamic variables; U-states defined by positions and momenta of individual molecules (Jaynes 1957).

c) M-states defined by mean and variance of incidence of nucleotide bases in the genetic code for a family of proteins; U-states defined by population of transition events that characterize spontaneous mutations (Holmquist and Cimino 1980).

d) M-states defined by distribution of forest areas killed by southern pine beetle; U-states defined in part by distribution of the population of beetles (Gold, Mawby and Hain 1981; Mawby and Gold, in review). The fact that the full nature of the U-space is unknown is relevant to the discussion of Subsection 2.3.

The problem is to make inferences concerning the microspace on the basis of information regarding the macrospace. In keeping with the underlying concept of the maximum entropy formalism, we wish to make such inferences on the basis of full use of information or data, but without imputation of "information" which we do not have. In a technical sense, we wish to determine a family of probability distributions on the microspace for which observations on the macrospace of the given type, comprise a minimal sufficient statistic. While this is the desired goal, we note that the problem stated in this form may not always have a solution.

We assume the existence of a measurable function g,

$$g: \ U \to M \ . \tag{1}$$

The mapping g may be defined by a model relating underlying variables to observables. Through such a map, the M-space induces a partitioning of the U-space; a given point or subset of the M-space is associated with its inverse image under the mapping g. This set in the U-space becomes the set of microstates allowable given the constraint of the specified (that is, observed) macrostate in the M-space. Given a point in the M-space and *no further inforation*, we are forced to assign equal likelihood to allowable points in the U-space; there seems to be little choice but to follow the Laplace rule of equiprobability.

Generalizing, a measure on U-space, $\gamma(u)$, together with a function g, defines a measure on the M-space according to

$$\nu(A) = \int_{g(u)\varepsilon A} d\,\gamma(u) \tag{2}$$

where $u \varepsilon U$, A is a subset of M-space, and the integral is over the region for which $g(u) \varepsilon A$. A particular probability density function p_U on U defines a probability function p_M on M by

$$p_M = \int_{g(u)\,\varepsilon A} p_U(u)\; d\gamma\,(u) \tag{3}$$

Our interest in equation (3) concerns the inverse problem: given p_M or at least some information about p_M, we wish to infer a probability distribution p_U on the U-space.

The simplest case, which we now consider, has the following characteristics: M-space is of finite dimension N; U-space is a finite state space with elements (states) $u_1, \ldots, u_n$; the function g is expressible as $g = (g_1, \ldots, g_N)$; the set A in equation (3) contains a single point $m = (m_1 = \ldots, m_N)$, with

$$m_i = E_{p_U}[g_i(u)],\ i = 1, \ldots, N, \tag{4}$$

where E_{p_U} denotes expectation with respect to a given distribution p_U .

In accordance with the goal set earlier, we seek that distribution p_U which carries the greatest degree of uncertainty regarding U, consistent with the requirements set by equations (4).

A scalar measure of the uncertainty associated with a probability distribution p is needed. The measure generally regarded as most natural is the entropy function S(p) introduced by Shannon (1948) as a measure of missing information,

$$S(p) = -\sum_{j=1}^{n} p_j \log p_j \tag{5}$$

where p_j is the probability assigned to the j^{th} allowable point in U-space. The choice of S(p) as a measure of the uncertainty rests on the demonstration (Shannon 1948) that it is the only expression (unique up to a positive multiplicative constant) with the following intuitively desirable properties:

1. S(p) is continuous with respect to the p_j .
2. For $p_j = n^{-1}$, $j = 1, \ldots, n$, S(p) is a monotonically increasing function of n.
3. S(p) is additive under decomposition of the space; that is, if $\{V_1, \ldots, V_k\}$ is a partition of the support set, the uncertainty of the distribution is equal to the uncertainty of the probability associated with membership in the subsets V_i plus the appropriately weighted uncertainties associated with the conditional distributions over points in the subsets V_i .

The formalism rationalized in this way, when it applies, leads to the constrained maximization problem

$$\max_{p_U} S(p_U) \tag{6}$$

subject to $\quad E_{p_U}[g_i(u)] = m_i$.

Such a problem is typically addressed using the method of Lagrange multipliers. Jaynes (1978) gives the solution in the form

$$p_j = \frac{1}{Z(\lambda)} e^{-\lambda \cdot g(u_j)} \tag{7}$$

where

$$Z(\lambda) = \sum_j e^{-\lambda \cdot g(u_j)} \tag{8}$$

$$\lambda_i = \frac{\partial S(p)}{\partial m_i} \quad . \tag{9}$$

As discussed by Jaynes (1968, 1978, 1982), the procedure yields that distribution which gives equal weight to all distributions consistent with the constraint equations. A variety of authors have rationalized the procedure from different points of view (for example, Akaike 1982, Christensen 1981; Shore and Johnson 1980; Tribus 1969). An especially compelling property of the problem as formulated in (6) is that it yields a family of distributions for which the functionals of equation (4) comprise a minimal sufficient statistic (North 1970).

Foremost among the applications of this formalism has been that of Jaynes (1957, 1968 and 1982) in the theory of statistical mechanics. Another active area of application has been in spectral analysis of time series, based on the work of Burg (1975). Virtually the only biological use of the formalism known to the author is that of Holmquist and Cimino (1980), indicated in example (c), above, which concerns probabilities of spontaneous mutation. It is important to point out that a large part of the *art* in the application of this methodology is in the correct formulation of the constraints (that is, choice of appropriate M-space) so as to make the problem computationally tractable.

Many important applications in biology are more complex in several respects. Before turning our attention to dynamic systems, we discuss three of the most important of these: noisy data; availability of prior information; incomplete characterization of the U-space and lack of a well-defined measurable mapping g from U to M.

2.1. The Problem of Noisy Data.

In terms of the basic problem, we introduce a space D, the data space, and a set W of measurement noises w. The diagram of the basic problem becomes

$$\begin{tikzcd} U \arrow[r, "g"] & M \arrow[r] & D \\ & W \arrow[ur] & \end{tikzcd}$$

That is, $m \varepsilon M$ is a function of $u \varepsilon U$, while $d \varepsilon D$ is a function of m and a random measurement noise $w \varepsilon W$. The result of the measurement process is therefore not a point in M-space, but a probability distribution on M-space, determined by d and by p_W, a given probability distribution on W (possibly a function of d). Such a problem is considered by Gull and Daniell in the context of applications in radio astronomy and pattern recognition (1978, also Daniell and Gull, 1980). They consider the special case of the data consisting of a collection of measurements on the components of $m = (m_1, \ldots, m_N)$ subject to independent Gaussian errors. In their procedure, $u \varepsilon U$ would be a particular point in the state space; i.e., a particular pattern of the sky (a particular distribution of light intensities). The probability of that pattern given the data is,

$$P(u|d) \propto P(u)\ P(d|u). \tag{10}$$

The object is to arrive at that u for which $P(u|d)$ is greatest, i.e., to maximize the posterior probability. In the case of independent Gaussian errors, $P(d|u)$ has the form $\exp(-X^2/2)$, where

$$X^2 = \sum_{i=1}^{N} \frac{(\hat{d}_i - d_i)^2}{\sigma_i^2} \tag{11}$$

$$\hat{d}_i = E_{p_W}(d_i|u)$$

and σ_i^2 has the usual meaning of variance associated with d_i. The quantity X^2 is distributed as χ^2 with N degrees of freedom. Gull and Daniell make this the basis of an iterative scheme. The prior is expressed as being proportional to exp[S(u)/z], where S(u) is the entropy associated with the distribution u, and z is a constant. Following equation (10), we are lead to maximize $S(u) - z[X(u)]^2$. At each iteration, we compute X(u) using the result of the previous iteration, and find a new value of u which maximizes $[S-zX^2]$. Jaynes (1982) gives a theoretical interpretation of the coefficient z, which suggests how its value should be chosen. However, the treatment does not yield an operational method and does not offer an alternative to the ad hoc method of Gull and Daniell, who choose z so as to cause $X(u)^2$ to have its expected value under the χ^2 distribution.

A problem arises because the data gathered in the applications of interest often have errors which are correlated and non-Gaussian. Indeed, a large portion of the statistical research literature deals with development of methodology that does not depend (or depends only weakly) on the assumption of non-correlated Gaussian errors. In the current context, we must be concerned with robustness of the result to specific types of violations of the assumption, and the possible need for extension of the formalism to other types of error structure.

2.2. Need for a Non-uniform Prior.

The information we wish to bring to bear is in many cases not contained in a single observation, even taking the observation to be vector valued. We may have a long history of past observations on the system of interest, as well as on related systems. The use of such information as additional constraints in a constrained maximization formalism in general presents insurmountable practical as well as theoretical difficulties. Nevertheless, acknowledging that the expressed constraint is incomplete requires that the Laplace rule of equiprobability be replaced by a more appropriate distribution over the set of U-states not proscribed by the constraint. Shore and Johnson (1980 and 1981) address this problem through the I-divergence measure of Kullback (1959), which generalizes the maximum entropy principle of Jaynes.

In an information sense, the distribution p which minimizes (subject to the given constraints) the expression,

$$\sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} , \tag{12}$$

where q = $(q_1, \ldots, q_n)$ is the prior, may be interpreted as the nearest distribution which satisfies the given constraints (Kullback 1959, Csiszar 1975). It follows that minimization of (12) gives that distribution p which, starting from distribution q, introduces the least amount of information consistent with the constraints. This prescription, which Shore and Johnson refer to as that of minimizing "cross-entropy," reduces to the maximum entropy rule when the prior is the equal likelihood distribution. Moreover, expression (12) generalizes to the functional

$$S(p,q) = \int p(u) \log \frac{p(u)}{q(u)} d\,\gamma(u), \tag{13}$$

which retains desired convergence and invariance properties. This is not true of the Shannon expression.

Shore and Johnson discuss computational algorithms based on minimization of expression (13). Of particular interest relative to the intended applications is the minimization of this expression relative to noisy constraints, which does not yet seem to have been addressed.

2.3. Stochastic Map from U to M.

It is most generally the case in biological science that the nature of our knowledge prevents specification of a well-defined measurable map from U to M-space. We conceive of the problem as arisingat least in part from the complexity (i.e., high dimensionality) of the U-space. The result is that we are not concerned with the full U-space, with elements $u = (u_1, \ldots, u_n)$ but with a projection onto a lower dimensional space, say $U^{(k)}$ of dimension k. That is, we write

$$u = (u^{(k)}, u^{(n-k)}) .$$

The associated point in M-space is determined by the sub-vector $u^{(k)}$, together with the neglected sub-vector $u^{(n-k)}$. The object then becomes to determine a marginal distribution on $U^{(k)}$ given a point (or possibly a distribution) on M-space. The mapping $U^{(k)} \twoheadrightarrow M$ becomes a stochastic map, which associates a point in $U^{(k)}$ with a probability distribution on M-space, rather than with a point in M-space.

An example is the relation between phenotype and genotype, which might be influenced by neglected environmental variables, so that a given genotype is associated with a distribution of phenotypes. Another example would be the dependence of bark beetle epidemic dynamics on beetle population and weather variables, which is influenced by a variety of other variables whose nature is not yet understood. Indeed this situation is the rule rather than the exception in biological investigations, and questions related to this point have been of central concern in the philosophy of biological science.

The point seems closely related to the question of noisy constraints. The mapping which determines d_i in equation (11) is now a composite of the measurement error which maps a point m into a distribution in D and the stochastic map which takes a point in $U^{(k)}$ to a distribution in M. In applying the algorithm of Gull and Daniell, the distributional difficulties (violation of assumption of non-correlated Gaussian errors) that arise from this source are likely to be more severe.

3. AN APPROACH TO EXTENDING THE FORMALISM TO DYNAMIC SYSTEMS.

Applications of both the maximum entropy and the minimum cross-entropy formalisms have been limited to estimation of a single time-independent distribution on U-space (see, however, Jaynes 1980). On the other hand, our primary interest is in making inferences concerning the dynamics of processes that transform a probability distribution on U-space to another distribution at a later time. In developing an approach to this problem, we borrow from the concepts of system morphism as developed by Zeigler (1976) and from the language of category theory as applied to system theory (Arbib and Manes 1975, Padulo and Arbib 1974, Chapter 8).

For the immediate discussion, we idealize to two points in time, t_1 and t_2, and restrict to the framework of the basic problem (noiseless constraints, uniform or equal likelihood prior, measurable function from U to M). We assume also that the M-space and U-space are time invariant as in the mapping $g: U \twoheadrightarrow M$. Our information may consist of a pair $(m(t_1), m(t_2))$ or a collection of such pairs assumed to be drawn from the same probability distribution on U-space and governed by the same dynamics. The relation is diagrammed as follows,

$$\begin{tikzcd}
m(t_1) \arrow[r, "F"] \arrow[d, "g^{-1}"] & m(t_2) \arrow[d, "g^{-1}"] \\
u(t_1) \arrow[r, "f"] & u(t_2)
\end{tikzcd}$$

Here, F is the observed (no measurement noise) relation between $m(t_1)$ and $m(t_2)$ and f is a stochastic transformation that carries $u(t_1)$ into $u(t_2)$. In a direct extension of the previous argument, we hypothesize that the pair or collection of pairs in $(M, t_1) \times (M, t_2)$, together with the Laplace principle of equiprobability or other appropriate prior, defines a joint probability distribution on the space $(U, t_1) \times (U, t_2)$ and the space of transformations f. If the space of these transformations is limited to a given parametric family (as may be required to insure measurability as well as tractability), the problem becomes one of inducing a probability distribution on $(U, t_1) \times (U, t_2) \times B$, where $B = \{b\}$ is the parameter space. In terms of the maximum entropy formalism, or the more general minimum cross-entropy formalism, we conjecture that the problem may be cast in terms of minimizing the cross-entropy of the joint distribution relative to a given prior and subject to the following types of constraints:

a) Constraintsof the type of equation (4), which would apply separately at each time, except that the distribution at either time would be conditional upon observations at the other time (assuming they are known at the time of inference).

b) Commutation of the diagram. In the current context, this would mean that for given functionals F, f and g, starting at the top left, then going to the right and down should give the same result as first going down and then to the right. In a deterministic framework, this implies the set identity,

$$f\{g^{-1}[m(t_1)], t_1, t_2, b\} = g^{-1}\{F[m(t_1)], t_1, t_2]\} \tag{14}$$

In the stochastic framework, equation (14) becomes an equality constraint between the probability distributions generated by the two sequences of operations. We further require that this commutation be invariant under change of variables.

In a relatively simple type of biological example, the components of the figure might have the following interpretations:

m: a measure of the size of a population;

u: a particular distribution of individuals over such variables as age, sex and space occupied;

F: the observed population level growth law;

f: the underlying dynamic that governs the reproductive and mortality characteristics of individuals.

In such a case, we might be especially interested in learning about the stochastic map represented by f.

Conceptually at least, it is straightforward to generalize to the case of multiple times (involving applications to filtering, prediction and stochastic control), and to the cases of noisy observations and stochastic mapping from U to M. Further useful generalizations might involve time dependent error structure on the observations.

ACKNOWLEDGEMENTS

I should like to thank John Burg and Ronald A. Howard (Stanford University) for helpful discussions on the concepts of maximum entropy and minimum cross-entropy, and H. R. van der Vaart and Marjolein V. Smith (NCSU) for constructive comments and critical review of an earlier draft.

REFERENCES

Akaike, H. (1982). Prediction and Entropy, MRC Technical Summary Report #2397, Math. Res. Center, University of Wisc., Madison.

Arbib, M. A. and E. G. Manes (1975). Arrows, Structures and Functors, Academic Press, New York, pp. 183.

Burg, J. P. (1975). Maximum Entropy Spectral Analysis, Ph.D. dissertation, Dept. of Geophysics, Stanford University, Stanford, California, pp. 123.

Christensen, R. A. (1981). Entropy Minimax Sourcebook, Entropy Ltd., Lincoln, Mass., pp. 575.

Csiszar, I. (1975). "I-Divergence Geometry of Probability Distributions and Minimization Problems," Annals of Probability 3:146-158.

Daniell, G. J. and S. F. Gull. (1980). "Maximum Entropy Algorithm Applied to Image Enhancement," Proc of IEE 127E:170-173.

Gold, H. J., W. D. Mawby and F. P. Hain. (1981). "A Modeling Hierarchy for Southern Pine Beetle," Proc. Symposium on Southern Pine Beetle Population Modeling, Asheville, North Carolina.

Gull, S. F. and G. J. Daniell. (1978). "Image Reconstruction from Incomplete and Noisy Data," Nature 272:686-690.

Holmquist, R. and J. B. Cimino (1980). "A General Method for Biological Inferences: Illustrated by the Estimation of Gene Nucleotide Transition Probabilities," BioSystems 12:1-22.

Jaynes, E. T. (1957). "Information Theory and Statistical Mechanics," Phys. Rev. 106:620-630.

Jaynes, E. T. (1968). "Prior Probabilities," IEEE Trans. Systems Sci. and Cybernetics, SSC-4:227-241.

Jaynes, E. T. (1978). "Where Do We Stand on Maximum Entropy" in The Maximum Entropy Formalism, eds., R. D. Levine and M. Tribus, MIT Press, Cambridge, Mass.

Jaynes, E. T. (1980). "The Minimum Entropy Production Principle," Ann. Rev. Phys. Chem. 31:579-601.

Jaynes, E. T. (1982). "On the Rationale of Maximum-Entropy Methods," Proc. IEEE 70:939-952.

Kullback, S. (1959). Information Theory and Statistics, Wiley, New York, reprinted 1978 by Peter Smith Publishers, Gloucester, Mass.

Mawby, W. D. and H. J. Gold (in review) "A Stochastic Model for Large Scale Southern Pine Beetle (Dendroctonus frontalis) infestation in the Southeastern United States."

Mesarovic, M. D., D. Macko and Y. Takahara. (1970). Theory of Hierarchical, Multilevel Systems. Academic Press, New York, pp. 294.

North, D. W. (1970). The Invariance Approach to the Probabilistic Encoding of Information, Ph.D. dissertation, Operations Research Department, Stanford University, Stanford, California.

Padulo, L. and M. A. Arbib. (1974). System Theory, Hemisphere Publishing Corporation, Washington, D. C.

Shannon, C. E. (1948). "The Mathematical Theory of Communication," Bell System Technical J., July-October, reprinted 1964, in The Mathematical Theory of Communication, by C. E. Shannon and W. Weaver, U. of Illinois Press, Urbana.

Shore, J. E. and R. W. Johnson. (1980). "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy," IEEE Trans. Information Theory IT-26:26-37.

Shore, J. E. and R. W. Johnson. (1981). "Properties of Cross-Entropy Minimization," IEEE Trans. Information Theory IT 27:472-482.

Tribus, M. (1969). Rational Descriptions, Decision and Designs, Pergamon Press, New York.

Zeigler, B. P. (1976). Theory of Modeling and Simulation, Wiley, New York, pp. 435.