July 1982

PRIOR INFORMATION IN INFERENCE*

E. T. Jaynes

Department of Physics, Washington University

St. Louis, Missouri 63130

Abstract: The Statistical Mechanics developed by physicists for predicting thermodynamic properties, also provides a formal mathematical procedure for incorporating prior information into general statistical inference. In some currently important problems, this can accelerate the process of finding appropriate models.

CONTENTS

* An Invited Essay written for the Journal of the American Statistical Ass'n.

INTRODUCTION

For many years the writer has carried an evangelical message to physicists, telling them: "Statisticians have learned many important things that physicists ought to know about but don't". Here we face in the opposite direction and note some things physicists have learned about inference, that statisticians might find useful.

A sociologist has complained that "God gave the easy problems to the physicists". While some of us would wish to qualify that statement, we shall add only: "--- and He so arranged that the solutions physicists found would help also in solving the problems of others".

What we have to say should be pertinent in some current problems, particularly those calling for a generalized inverse (i.e., given the data vector $d = Ax + e$ where A is a singular operator, estimate the "state of Nature" vector x). Here sampling theory finds itself groping for a defensible algorithm, because the data cannot distinguish states x, x' differing by a solution of the homogeneous equation $A(x-x') = 0$. Collinearity in multiple regression, missing cells in ANOVA, and power spectrum estimation from incomplete autocovariance data, are examples. Yet what the data cannot distinguish, the prior information often can.

In conventional sampling theory the act of choosing a model is just an intuitive way of taking account of certain prior information about the problem. This process can be, at least in part, removed from the realm of intuition and reduced to a formal mathematical procedure. In some cases prior information can create our model and parameters for us, out of the data, in a way that is optimal by a certain well-defined criterion (this "adjunct model" makes the best use of the data).

These things have been, in a sense, "well known" to physicists for decades, but in a language and context so different that they were not recognized. To explain, without going into lengthy details, why this happened and how the connection was finally made, we turn to a fable somewhat in the spirit of Fisher's "Problem of the Nile".

A MODERN FABLE

Once upon a time, at a mythical Agricultural Experiment Station, workers faced problems of inference in which they had a great deal of noise, but very little prior information. Being rational people, they developed methods--one of which was called maximum likelihood--that dealt most effectively with these problems. That is, they represented the properties of the noise by what they called a "sampling distribution" and did not bother with prior information that would have made little difference in their conclusions. Being also human, they were pleased at the success of their efforts, and made an inductive generalization of the kind everybody makes: since these methods have been successful in the problems we have studied, they must be the correct methods for all problems of inference. Their descendants formed what we shall call Camp A.

At an equally mythical Physics Laboratory, workers studied problems of inference related to thermodynamics; for example, given the energy and volume of a system composed of billions of molecules, predict its pressure, specific heat, etc. In these problems they had a great deal of prior information (all the known laws of physics as applied to the individual molecules of the system) but very little noise (their measuring instruments were quite good). Being rational people, they developed methods--one of which was called maximum entropy--that dealt most effectively with these problems. That is, they concentrated on finding the distribution, which

they called an "ensemble", that represented their prior information as constrained by the data; and did not bother with noise. They too were pleased at the success of their efforts and drew the obvious inductive generalization: henceforth, all problems of inference must reduce to finding the ensemble of maximum entropy. Their descendants formed Camp B.

Up to this point, our fable is not new. It is just the old story of the elephant and the blind men, each of whom thought the whole elephant must be like the part of it that he had touched. But there is a sequel:

Now there was a wise man who lived between these camps, observing them. Like Socrates, he made himself highly unpopular with both by asking thought-provoking questions. To the A camp he said, "Why do you not take prior information into account? See how easy it would be--you have only to introduce a prior probability distribution and add the log prior to your log likelihood before maximizing." To the B camp he said, "Why do you not take noise into account? See how easy it would be--you have only to introduce a sampling distribution and add the log likelihood to your log prior before maximizing."

But the Tower of Babel Syndrome had broken out and there was a language difficulty. Those in camp B had never used the term "prior probability" because for them an ensemble was not, logically or chronologically, "prior" to anything else; it was the only distribution in sight. The "log prior" was, essentially, what they called "entropy". So they did not understand what the wise man was saying.

For those in camp A, the very word "probability" had come to have a different meaning than the original one. As a result, they now believed that the equations of probability theory were only rules for calculating

frequencies; and not for conducting inference. So they thought the wise man was asking them to compromise the "objectivity" of their inferences with personal opinions, unsupported by any data. They reacted with high-minded scorn and indignation and coined the slogan "Let the data speak for themselves!"

The wise man's suggestion therefore had no effect on the actual practice of either camp. Indeed, the situation deteriorated: instead of accepting his advice, which would have moved them together into a single camp, the two camps went their separate ways, developing not only totally different languages, but also totally different conceptual foundations, that had the effect of institutionalizing their differences instead of resolving them.

Now in the fullness of time it came to pass that new problems appeared, bearing such curious names as "time series analysis" and "image reconstruction", in which the noise and the prior information were equally important. Each camp moved in, filled with confidence from past successes. Since each had a method that dealt properly with half the problem, each was able to extract about half of the correct solution; and that much is often good enough to be usable. But of course their solutions were quite different.

In estimating the power spectrum of a time series, given the same data both would find the strong lines in about the right positions, but with different shape and resolution. Solution A, considering noise but not prior information, achieved good stability with respect to noise, but suffered from poor resolution and spurious "side-lobe" responses. Solution B, considering prior information but not noise, achieved very high resolution without side-lobes, but suffered from variability and spurious responses to noise.

Here, so to speak, the whole elephant finally came into view. But because of their language and conceptual differences, each camp found the rationale of the other's method incomprehensible. Each saw in the different results, not evidence of the incompleteness of his own method, but a proof of the defects in the other's method.

And so our story ends in stalemate, the Tower of Babel Syndrome in full control. Unable to communicate in a common language, each camp has the solution to the other's problem, and cannot recognize that the other has the solution to its own problem.

APOLOGY

A fable is, by definition, something that is not literally true, but which is thought by its writer to convey a true moral. In the reality, differences in technical detail and chronology made the situation more complicated than is portrayed above. Yet some 30 years' immersion in the technical details of both fields has led the writer to this rather laconic view of the relation between two methods of inference, generally called "sampling theory" and "Gibbsian statistical mechanics", and developed respectively by statisticians and physicists.

But not all statisticians are sampling theorists, nor all physicists Gibbsians. The neutral terms "camp A" and "camp B" indicate our basic concern with two different approaches to inference, not two different professions.

Our wise man is a composite of J. Bertrand (1889), H. Poincaré (1912), Sir Harold Jeffreys (1939), I. J. Good (1950), and L. J. Savage (1954).

I still recall with a shudder the scorn and indignation of a well-known Professor of Statistics when, as a student at Princeton in the late 1940's (with John Tukey and Paul Meier among my friends) I asked, naively,

why he did not take Jeffreys' advice and improve his confidence intervals by incorporating prior information into them. In camp A it was considered to be, not just illogical, but a morally reprehensible breach of "scientific objectivity" to allow one's self to be influenced by prior information. This scruple would undo all the useful results found in camp B.

Yet this vast difference in attitude in the two camps does not, in the writer's view, reflect any difference in basic thought processes. It is only an historical accident arising from the very different problems we encountered in our formative years; and the situation might have been the other way around. That is, Gibbsian statistical mechanics is a rather late development in physics, while Rothamsted occurred at a relatively early stage in the development of biology.

If the Rothamsted workers had been in possession of the exact "biochemical equations of motion" telling how each individual cell of a growing plant responds to its environment, it would have been obvious that this prior information must be taken into account in estimating differences in yield from two varieties or two treatments. Likewise, if physicists had not possessed such good measuring instruments, it would have been obvious that thermodynamics must take noise into account.

Recognizing this, could we not develop an area of common language that emphasizes the basic unity of all inference? As a start, we try to explain the rationale of Gibbsian statistical mechanics in a language closer to that of statisticians, but without lapsing into statisticians' jargon that would be unintelligible to most physicists.

LOGIC OF STATISTICAL MECHANICS

This is, in essence, no different from the game of "twenty questions". We have an initial hypothesis space $H_0 = (h_1 \ldots h_n)$ determined by our prior information about the nature of atoms. H_0 is enormous, comprising perhaps $n = \exp(10^{26})$ conceivable quantum states. Each new piece of information (data) that we acquire is a constraint that narrows down these possibilities to some $H_1 \subset H_0$.

That is really all there is to it. Difficulties in understanding this rationale come from the fact that it is usually described in a physicist's jargon that refers, not to the simple underlying idea, but to elements of the rather indirect mathematical formalism that has evolved to carry it out.

Certainty is never reached, and so whatever data we have managed to obtain, we must be prepared to answer at each stage: What are the best predictions we can now make, of the quantities of interest $A = (a_1, a_2, \ldots)$? If we later acquire more data, then our hypothesis space will (unless the data are redundant) be further contracted to $H_2 \subset H_1$ and in principle we shall of course expect to improve on the old predictions.

In practice, as our sociologist noted, physicists are lucky and we quickly reach a "plateau" stage where our predictions have become so good that further data hardly matters. That is, H_1 is already so homogeneous in A that further contractions are unnecessary. It is not the amount of contraction, but its homogeneity in A, that makes our predictions reliable. So the strategy is: what kind of data $D = (d_1, d_2, \ldots)$ should we seek, to get us to that plateau as quickly as possible? Stated differently: what observable quantities D are most strongly correlated, in our initial hypothesis space H_0, with the things we want to predict?

When a physicist looks for the "laws of thermodynamics", i.e., the reproducible connections between pressure, temperature, magnetization, specific heat, etc., his rationale is no different from that of an economist who looks for the most reliable indicators. Although the superficial appearances are at present entirely different, we are all doing basically the same reasoning, and all would benefit from recognizing this.

This principle of homogeneous contraction of the hypothesis space is, of course, just the rationale anybody does adopt naturally in his everyday problems of inference. A chemist trying to determine what elements are present in a sample, a medical diagnostician, a TV repairman, a burglar casing a new neighborhood--all start with a large class of conceivable hypotheses which they try to narrow down, as quickly as possible, by acquiring data that are as relevant as possible to their various goals.

It is interesting to note that this reasoning is conducted without mention of any sampling distribution. Conversely, it is quite foreign to the outlook of sampling theory to think of data as constraints on a prior hypothesis space. This illustrates the completely different language and conceptual foundations that have developed in camp A and camp B.

HOW SHOULD WE USE PRIOR INFORMATION?

In most real-life situations no sampling distribution is given to us by Nature, and we have no general principles, that could be taught in statistics courses, to determine what our initial hypothesis space H_0 should be. This is, necessarily, a matter of judgment based on knowledge of the subject-matter. Clearly, our search for the right hypothesis will be shortest if every bit of prior information, that would help to restrict that space, is taken into account from the start.

In spite of their disdain for prior probabilities, sampling theorists do recognize the relevance of such prior information; and as noted, they use it in deciding on the initial formulation of a problem. Presumably, no rational person would define a parameter space Θ that includes values of θ known *a priori* to be impossible; or excludes values that seem *a priori* possible. But if we stop at that point, we have missed something very important. What is it that makes the physicist's contracted hypothesis space so homogeneous with so little data?

Most physicists and many engineers are without formal training in statistics, and turn to the orthodox statistical literature only for help with some specific application. But they perceive instinctively, if sometimes vaguely, that merely specifying a sampling distribution and parameter space cannot represent all the knowledge they have, that is relevant for their inferences. It is typical of real scientific problems that one has some kind of direct, highly cogent prior information about the likely values of θ that has nothing to do with frequencies in any "random experiment". Most, finding in the current literature no way of using this information, devise their own *ad hoc* procedures that rely on judgment rather than orthodox statistical theory.

Indeed, it is not only physicists and engineers who have perceived this. The necessity of incorporating prior information into the actual process of inference--and not merely using it in deciding on the initial formulation of a problem--was noted by J. Bertrand (1889) the year before R. A. Fisher was born. Of Bertrand's several examples, we quote the last:

"The inhabitants of St. Malo [a small French town on the English channel] are convinced; for a century, in their village, the number of deaths at the time of high tide has been greater than at low tide. We admit the fact.

"On the Coast of the English channel there have been more shipwrecks when the wind was from the northwest than for any other direction. The number of instances being supposed the same and equally reliably reported, still one will not draw the same conclusions.

"While we would be led to accept as a certainty the influence of the wind on shipwrecks, common sense demands more evidence before considering it even plausible that the tide influences the last hour of the Malouins.

"The problems, again, are identical; the impossibility of accepting the same conclusions shows the necessity of taking into account the prior probability of the cause."

Clearly, Bertrand cannot be counted among those who advocate "letting the data speak for themselves". Such adages as "Correlation does not imply causation" or "An empirical fit is no substitute for a reason" also recognize that the data are only the latest addition to our knowledge, not the whole of it. But we must become more specific: what is the quantitative factor missing from camp A calculations but supplied by camp B?

THE MISSING MULTIPLICITY FACTOR W

The point we wish to make was recognized in a famous discussion between d'Alembert and Laplace. Bernoulli had given a rule for assigning probabilities: P(A) = (number of cases favorable to A)/(total number of equally possible cases). Now predict the result of tossing two coins:

d'Alembert: "There are three possibilities: (Two heads), (One head, one tail), (Two tails). Therefore, by Bernoulli's rule we assign probabilities (1/3,1/3,1/3)."

Laplace: "No!! There are four possibilities: (HH, HT, TH, TT) so we should assign probabilities (1/4,1/4,1/4,1/4). The event (One head, one tail) is more likely because it can happen in two different ways (i.e., it has a multiplicity W = 2) and we ought to take that into account. d'Alembert should have assigned (1/4,1/2,1/4)."

For 200 years all schools of thought--whether or not they accept Bernoulli's rule--have agreed that the probability of an event A is the sum $P(A) = \sum P(a_i)$ over the mutually exclusive ways in which A can occur. If the $P(a_i)$ are all equal, the multiplicity factor W appears automatically.

But these things are well known to everybody; the central limit theorem describes how multiplicity factors pile up into a gaussian when convolved many times. Indeed, the derivations of sampling distributions that launched Fisher's career were, in essence, ingenious ways of reasoning out the multiplicity factors for various functions z(D) of the data D. How then can one say that sampling theory ignores multiplicity?

Well, sampling theory does take correct account of multiplicity in the sample space; the trouble is that a sampling distribution $P(D|\theta)$ says nothing about the multiplicity of the parameter space Θ. Yet our prior knowledge of the phenomenon being observed might tell us that θ has a definite, calculable multiplicity $W(\theta)$.

Any method of inference about θ which looks only at sampling distributions conditional on θ may be missing something of crucial importance for the inference. If we know that the value $\theta = 0$ can occur in only one way, while $\theta = 1$ can happen in 10 different ways, there are 100 different ways in which Nature can generate the value $\theta = 2$, etc., that is information that we shall ignore at our peril in making predictions involving θ.

EXAMPLE: HOW DO PHYSICAL CHEMISTS DO IT?

In the laboratory one measures certain macroscopic quantities, such as temperature T, pressure P, etc.; and in the theory one tries to predict, from such data, other macroscopic properties such as total energy E, density R, magnetization M, heat capacity C, etc. Everybody accepts the rule found by Ludwig Boltzmann, that at temperature T the probability of a state of energy E is proportional to $\exp(-E/kT)$, where k ($= 1.38 \times 10^{-16}$ ergs/degree) is Boltzmann's constant, a corrective fudge factor necessitated by our curious system of units.

Then, given the temperature of this room, T = 290 K, what values of (R,E,M) for the air in it do we expect to observe? By the Boltzmann law, it appears that the state of lowest energy will be overwhelmingly the most probable, because k is so small. Thus all the air in the room where I am writing this should be condensed into a small frozen puddle at the lowest point of the floor. Evidently, the prediction has ignored something of crucial importance for the inference.

J. Willard Gibbs (1875, 1902), showed how to correct this. We introduce a new quantity, a function S(R,E,M) of the macroscopic state and define the "free energy" F = E-TS. Then we modify Boltzmann's rule by taking the probability of the state (R,E,M) proportional to $\exp(-F/kT)$.

Given T, the most probable state is the one that minimizes F. We now find that the corrected rule gives us, at all temperatures and to the accuracy of our best measurements, a quantitatively correct prediction of the condition of the air in this room. The new quantity S that has rescued us from being eternally frozen and immobile is called entropy.

The rule for constructing the entropy function $S(X_1,\ldots,X_n)$ for any macroscopic quantities $\{X_1,\ldots,X_n\}$ that we can measure experimentally was given by Gibbs for the case of thermal equilibrium states. One can apply it usefully, cookbook style, without understanding what entropy really is. But generalization to other problems requires some understanding.

Let $W(X_1,\ldots,X_n)dX_1\ldots dX_n$ be the number of quantum states of our system, for which X_i is in the range dX_i, $1 \leq i \leq n$. That is, W is the multiplicity of the state $\{X_1,\ldots,X_n\}$. If we make, following Boltzmann, Einstein, and Planck, the interpretation of entropy:

$$S = k \log W \tag{1}$$

the successful rule becomes: the probability of the macroscopic state $\{X_1,\ldots,X_n\}$ is proportional to

$$\exp(-F/kT) = W \exp(-E/kT) \tag{2}$$

and suddenly all is clear!

When the chemist replaces the total energy by the free energy, he is simply taking into account the missing multiplicity factor W, the number of ways (number of microscopic quantum states) in which the observable macroscopic state can be realized. Doing this converts a disastrously wrong inference into a reliable, quantitatively correct one. Unfrozen quantum states exist, not because any one is more likely than the frozen

one, but because there are so many more of them. For every way in which the air in this room can be frozen, there are something like

$$W \simeq \exp(10^{25}) \tag{3}$$

ways in which it can be unfrozen. An increase in energy too small to measure, may still correspond to an increase in multiplicity by a factor of $\exp(10^{10})$.

This shows why multiplicity factors forced themselves on our attention in thermodynamics, before quantum states were discovered. In the real world, as Max Planck put it, Nature will appear to have a "strong preference" for those situations of highest entropy. This preference was discovered experimentally in the first half of the nineteenth century, long before its explanation was found; and it was called the "Second Law of Thermodynamics".

Boltzmann called W the "thermodynamic probability", an unfortunate terminology because multiplicity is, like likelihood, not a true probability but only one of the factors in a probability. The resulting preference of Nature was expressed in the variational principle of Gibbs (1875): all thermal equilibrium states can be predicted quantitatively by maximizing the total entropy of (system + environment) subject to the constraints operating, both those imposed by the experimenter and those arising from laws of physics (conservation of mass, energy, number of atoms, etc.). For a century, physical chemistry has been based on this principle.

Similarly, if $W(\theta)$ varies over a parameter space by a factor of only 10 or 100, maximum-likelihood estimates would still have tolerably good success. But when $W(\theta)$ varies by many orders of magnitude in a small interval of θ, we can hardly expect to make even qualitatively right inferences about θ if that fact is ignored. Multiplicity is equally essential for predicting

a "parameter" or a "random variable", and so a sampling theorist will be most successful if his intuition leads him to define his parameters so that, like location parameters, they have nearly uniform multiplicity.

RELATION TO PRIOR PROBABILITIES

Multiplicity is one factor--often the only variable factor--in a prior probability distribution. Often in the past, prior probabilities have been rejected as vague and ill-defined. Today, it would be misleading to repeat such criticisms without taking note of the progress made on these problems in the recent Bayesian literature. In any event, there are many important applications where our prior information makes the multiplicity $W(\theta)$ a very well-defined quantity, far more "objectively real" than are those "iid normal" sampling distributions assigned merely by convention. Calculation of $W(\theta)$ may be a nontrivial combinatorial problem.

Consider the simple case where the only combinatorial result needed is the multinomial coefficient. Nature generates a set of non-negative integers $\{N_1 \ldots N_n\}$ which we may represent by the fractions $f_i = N_i/N$, where $N = \Sigma N_i$. We have some data D and a sampling distribution $P(D|f_1 \ldots f_n)$, from which we are to estimate some function $G(f_1 \ldots f_n)$.

The new feature, which makes this different from a conventional sampling theory problem, is that we have prior information about the process that generated the set $F \equiv \{f_1 \ldots f_n\}$, so we know that the multiplicity of F (number of ways in which it could have been realized) is

$$W(F) = N!/[(Nf_1)!(Nf_2)! \ldots (Nf_n)!] \quad . \tag{4}$$

Knowledge of W(F) can affect our inferences, just as it did in the case of the physical chemist. In decision theory, F is called the "state of nature";

wanting a shorter term and with a view to applications in image reconstruction, we call it here simply the "scene". Analogous to (1), we shall say that any scene F possesses an entropy log W(F). Given (N,n), the number K of different possible scenes is equal to the number of terms in the multinomial expansion of $(f_1 + \ldots + f_n)^N$:

$$K = (N+n-1)!/N!(n-1)! \quad . \tag{5}$$

In physicist's jargon this is "the multiplicity factor for Bose-Einstein statistics".

We examine first the "no noise" problem, where the contrast with pure sampling theory methods stands out most clearly, after which we shall take the advice of that wise man and combine the two methods into a single procedure.

THE NO-NOISE PROBLEM

By "no noise" we mean that there is no sampling distribution except in the rudimentary sense that $P(D|F) = 1$ or 0; any data set D either is or is not the one generated by the scene F. The problem is that many different scenes all generate the same data and therefore have the same likelihood:

$$D = AF \tag{6}$$

where A is some operator, by hypothesis known but not uniquely invertible. Any rule for estimating F from D is, symbolically,

$$\hat{F} = RD \tag{7}$$

where R is a "resolvent" operator to be chosen.

It would appear that any rational choice of R must have, at the very minimum, the property that $\hat{F}$ lies in the class C of logically possible scenes: for all F, $D = AF = A\hat{F} = ARAF$. Thus R must be a generalized inverse:

$$ARA = A \quad . \tag{8}$$

Stated differently, as seen through the "window" A, the estimated scene must be indistinguishable from the true scene, otherwise we have not used all our information. However trivial and obvious this requirement may seem, solutions that violate it (for example, power spectrum estimators that can become negative) have been advocated repeatedly in the literature. Indeed, the only thing the data can tell us about the true scene is that it lies in class C. The data provide no basis for choice within C.

Among current real problems encompassed by this model we note:

(I) Generalized Loaded Dice. A random experiment has n possible results at each trial, so in N trials there are in all $n^N = \Sigma W(F)$ conceivable outcomes, where N_i are the sample numbers, and $F \equiv \{f_1 \ldots f_n\}$ is the set of frequencies generated in a particular realization. Given some data D consisting of m functions $d_k(f_1 \ldots f_n)$, where $1 \leq k \leq m < n$, estimate the $\{f_i\}$ or decide whether there is evidence for a systematic deviation of a function $G(f_1 \ldots f_n)$ from its "null hypothesis" value G_0. An entropy analysis of the famous dice data of R. Wolf from this standpoint which led to definite conclusions about the imperfections of the die, was given recently (Jaynes, 1978).

(II) Image Reconstruction. N elements of luminance have been distributed over n pixels, to generate the scene F. Our data consist of m numbers, $D = \{d_1 \ldots d_m\}$, $m < n$, which constitute a blurred image of the true scene:

$$d_k = \sum_{i=1}^{n} A_{ki} f_i \ , \quad 1 \leq k \leq m \tag{9}$$

where the matrix A is the digitized point-spread function of our telescope. Out of the large class of scenes compatible with our data, which is most likely to be the true one? Recent image reconstructions based on entropy have been given by Frieden (1980) and Gull and Daniell (1978).

(III) Statistical Mechanics. N molecules have n possible quantum states each, the fraction in state i being $f_i = N_i/N$, $1 \leq i \leq n$. The "scenes" are now the possible distributions $F \equiv \{f_1 \ldots f_n\}$. A molecule in state i has energy E_i and magnetization M_i. Given the total energy, $E = N \Sigma E_i f_i$, predict the total magnetization $M = N \Sigma M_i f_i$. The literature of statistical mechanics throughout this century has dealt with hundreds of problems with this logical structure (although usually more complicated in detail). The success of the predictions that flow from the multiplicity W(F) is taken for granted just as confidently as for Newton's laws of motion.

(IV) Spectrum Analysis. Given the value of the autocovariance of a time series for the first (m+1) lags:

$$R_k = \frac{1}{N+1} \sum_{i=0}^{N} y_i y_{i+k} , \qquad 0 \leq k \leq m \tag{9}$$

estimate its power spectrum S(f). Many impressive results in problems of this type arising in geophysics have been achieved recently by Burg (1975), Currie (1980), and others by taking multiplicity into account.

From the standpoint of conventional sampling theory, each of these problems is grossly ill-posed and has no solution, for the likelihood is constant on C. Yet in each case, recognition of the multiplicity factor supplies the missing criterion of choice within C, and leads to solutions that are unique, calculable, useful, and empirically verifiable. To understand this success, let us calculate some multiplicities.

SOME NUMBERS

Suppose there are two scenes, 1 and 2, equally compatible with our data: $p(D|F_1) = p(D|F_2)$. The writer knows of no principle in sampling theory by which we could choose one over the other. But calculating

their multiplicities, we find that $W_2/W_1 = 10^{10}$. That is, for every way in which Nature could have generated scene 1, there are 10^{10} ways (about four times the number of minutes since the Great Pyramid was built) in which she could have made scene 2. I shall, rather confidently, place my bets on scene 2 as being the right one.

Someone will surely object here that I have made an unstated and unwarranted assumption: that all these ways are equally likely. To this there are two replies. First, in view of the factor 10^{10}, I need not "assume" very much. Indeed, unless I had prior information that Nature has, for other reasons, some strong predilection for scene 1 over scene 2--and by more than a factor of 10^{10}--my decision could not be changed. But of course, if I did have that kind of prior information, to ignore it would be an even greater sin than ignoring multiplicity. The objection deserves a more aggressive reply; we return to it in our concluding remarks.

But do we really have such enormous variations in multiplicity in the real problems that arise in image reconstruction and time series analysis? Let us get some idea of the numbers of involved in a real case. Gull and Daniell (1978) gave some beautiful examples of maximum-entropy image reconstruction in radio and x-ray astronomy. In most cases they used a 128 x 128 grid, thus generating an image of $n = 16384$ pixels (requiring 3 minutes on an IBM 370/165). If we suppose, rather conservatively, that they could discern a change in intensity of a pixel amounting to 10% of the average intensity, we have effectively $N = 10n = 163{,}840$ elements of luminance comprising the scene. If their intensity resolution was better than this, the following numbers are underestimates.

These values of N and n could make, according to (5),

$$\frac{180223!}{163840!\ 16383!} = 3 \times 10^{23840} \tag{10}$$

distinguishable scenes, whose multiplicities range from $W_{min} = 1$ to

$$W_{max} = \frac{163840!}{(10!)^{16384}} = 4 \times 10^{675703} . \tag{11}$$

Therefore, we could partition the conceivable scenes into 67571 categories according to their multiplicity, those in category c having multiplicity in the range

$$10^{10c} \leq W < 10^{10(c+1)} , \quad 0 \leq c \leq 67570 . \tag{12}$$

A higher value of c denotes a smoother scene; but the eye surely does not distinguish 30000 different degrees of smoothness. Thus two scenes, as alike as possible except that one lies in category c, the other in c + 2, will be virtually indistinguishable to the eye. Yet every scene in c + 2 has a multiplicity greater than any scene in c by a factor of more than 10^{10}.

In realistic problems, then, we not only have variations in multiplicity by a factor 10^{10} between two scenes, we have chains of thousands of such comparisons with a factor of 10^{10} at each step. As large numbers go, these are hardly in a class with those we encounter in statistical mechanics; yet by ordinary standards the numbers are rather respectable, and it seems evident that such multiplicity factors need to be taken into account in image reconstruction by the same kind of reasoning that is used in statistical mechanics.

That is, having seen these numbers, we expect that a kind of "second law of thermodynamics" will manifest itself, and Nature will appear to have a strong preference for those scenes which have highest multiplicity (entropy)

compatible with our data. The success of the maximum entropy reconstructions of Gull and Daniell is in no way surprising as soon as it is recognized, as an elementary combinatorial theorem, that with a plausible hypothesis about how Nature is forming the scene, the overwhelming majority of all possible scenes compatible with their data had entropy very close to the maximum. Conversely, if that hypothesis were significantly wrong, the failure of those reconstructions would tell us so, and give us a clue pointing to a better hypothesis. As the writer sees it, this kind of logic is the essence of the scientific method.

THE MAXENT FORMALISM

The above indicates how, in generalized inverse problems, entropy can provide the missing criterion of choice within the class C of possibilities allowed by our data. In the noiseless case all scenes in C have the same "likelihood" in the technical sense of that word, and are thus equally good from the standpoint of sampling theory. Yet in some cases the multiplicities vary over C by large numerical factors, making scenes of high entropy far more "likely" than others, in the colloquial sense of the word.

This reasoning is evidently quite general, having no necessary connection with thermodynamics or physics. Thermodynamics was, historically, the first application where these things were recognized.

The mathematical formalism by which one locates the point of maximum entropy was given, as a special case, by Boltzmann (1877); and in greater generality by Gibbs (1902). However, full appreciation of the power of the method outside thermodynamics has been achieved only quite recently, as a result of the development of computer programs capable of dealing with dozens to thousands of simultaneous constraints. We recall the MAXENT formalism

briefly, in a notation that frees it from its historical origins in thermodynamics, but without becoming so abstract that it no longer suggests any applications at all.

A real variable x may take on the discrete values $(x_1 \ldots x_n)$. In N trials these are realized with frequencies $(f_1 \ldots f_n)$ respectively; this defines the "scene" F. But the f_i are not observable directly; the available data D are incomplete, consisting of the m measurements, $m < n$:

$$d_k = \sum_{i=1}^{n} A_{ki} f_i \quad , \quad 1 \leq k \leq m \tag{13}$$

where A is a known but singular matrix. Equation (13) is a collection of m simultaneous constraints defining our class C of possible scenes. Required: to find the scene $\hat{F} = (\hat{f}_1 \ldots \hat{f}_n)$ which has maximum entropy log W subject to the constraints (13) and normalization $\Sigma f_i = 1$.

In the expression (4) for W we may use the Stirling approximation, since the $N_i = Nf_i$ are large. Then as $N \to \infty$ we have

$$H(F) \equiv \lim \frac{1}{N} \log W(F) = - \sum_{i=1}^{n} f_i \log f_i \tag{14}$$

the same expression as found by Shannon (1948), by an entirely different argument, as a fundamental "information measure". Analytically, it is much easier to maximize H(F).

The conventional method introduces a Lagrange multiplier λ_k for each datum d_k, $(1 \leq k \leq m)$ and yields the solution

$$\hat{f}_i = \frac{\exp\left(-\Sigma_k \lambda_k A_{ki}\right)}{Z(\lambda_1 \ldots \lambda_m)} \quad , \; 1 \leq i \leq n \tag{15}$$

where

$$Z(\lambda_1 \cdots \lambda_m) \equiv \sum_{i=1}^{n} \exp\left(- \sum_{k=1}^{m} \lambda_k A_{ki} \right) \tag{16}$$

is a basic generating function of the kind that arises in so many combinatorial problems, called by physicists the partition function. Then, as usual, the Lagrange multipliers are found by requiring (15) to satisfy the constraints (13). The result may be written as

$$d_k = - \frac{\partial}{\partial \lambda_k} \log Z \quad , \qquad 1 \leq k \leq m \tag{17}$$

a set of m simultaneous equations for the m unknowns $(\lambda_1 \cdots \lambda_m)$.

Of course, there is a great deal more detail in the full MAXENT formalism, a mass of covariance—reprocity—perturbation theorems and generalizations to continuous distributions and to quantum theory (Jaynes, 1978, 1980); but the above bare skeleton will suffice to indicate how most of the calculations are carried out. Equation (15) defines a "generalized Gibbsian canonical ensemble". By the Shannon interpretation it is the "most honest" representation of our knowledge of the true scene, when the only information about it consists of the data (13). That is, any other distribution would necessarily either assume information that we do not have, or contradict information that we do have.

For some 60 years, virtually all analytical calculations in statistical mechanics have started with the determination--exact or approximate--of the appropriate partition function Z. Once log Z is known, in its dependence on the Lagrange multipliers λ_k, essentially all physical predictions of interest follow, and the λ_k acquire various physical meanings. For example, if we define A_{1i} as energy of the i'th quantum state:

$$A_{1i} = E_i \quad , \qquad 1 \leq i \leq n \tag{18}$$

then the Lagrange multiplier λ_1 turns out to have the physical meaning $\lambda_1 = (kT)^{-1}$, and so the Boltzmann distribution $\exp(-E/kT)$ is a special case of (15).

In the above we have supposed the data noiseless. Suppose that instead of (13) we have $d_k = \Sigma A_{ki} f_i + e_k$, with $e_k \sim N(0,\sigma_k)$, the traditional gaussian noise. The wise man said: add the log likelihood to your log prior before maximizing. From Eq. (14), we shall then maximize $NH(F) + Q(F)$, where $Q(F) = -\frac{1}{2} \Sigma_k (d_k - \Sigma_i A_{ki} f_i)^2 / 2\sigma_k^2$ is the traditional quadratic form. The resulting reconstructed scene is the peak of a posterior distribution. This is just the way Gull and Daniell (1978) allowed for noise, and they show how the quality of a reconstruction varies with the noise level. Further details are in the process of publication, and should be available by late 1983.

Although there is much more to be said about the generalized inverse problem, let us turn now to some more general uses for (15). We have obtained it by a combinatorial argument, as an estimate of the frequency distribution that generated the data. It is the "best" estimate in the sense that, of all distributions consistent with those data, (15) has the greatest multiplicity. But the analytical result has other valuable properties beyond the generalized inverse area.

REINTERPRETATION - ADJUNCT MODELS

The relations to be noted next are mathematically trivial but conceptually subtle; we are trying to translate results long known in camp B into camp A language, although the basic ideas that led to them are not in camp A vocabulary and concepts. So the following is not a derivation, but only a line of free association. Given an estimate $\hat{f}_i$ of frequency such as (15), it seems reasonable even to one who does not define a probability as a frequency,

to assign $p(x_i) = \hat{f}_i$ for purposes of future prediction. In camp A language, having found the distribution (15), nothing prevents us from re-interpreting it as a sampling distribution, with parameters $\{\lambda_1 \ldots \lambda_m\}$.

As soon as we know the matrix A that defines the "nature" of the data, and before we have the data, we know that this distribution is going to have the analytical form

$$p_i = Z^{-1} \exp\left(- \Sigma_k \lambda_k A_{ki}\right) \quad , \quad 1 \leq i \leq n \quad . \tag{19}$$

Given the distribution (19), our "best" (by mean-square error criterion) prediction of the data would be

$$\hat{d}_k = E(A_{ki}) = \sum_{i=1}^{n} A_{ki} p_i \quad , \quad 1 \leq k \leq m \quad . \tag{20}$$

But this is

$$\hat{d}_k = - \frac{\partial}{\partial \lambda_k} \log Z \quad , \quad 1 \leq k \leq m \tag{21}$$

i.e., just Eq. (17), which now has a second meaning.

Finally, let us ask: given the sampling distribution (19) and the results of N trials in which x_i was obtained N_i times, what are the maximum-likelihood estimates of the parameters $\{\lambda_1 \ldots \lambda_m\}$? The log-likelihood is

$$L(\lambda_1 \ldots \lambda_m) = \sum_{i=1}^{n} N_i \log p_i \quad ,$$

$$= -N \log Z(\lambda_1 \ldots \lambda_m) - \sum_{ik} \lambda_k A_{ki} N_i \quad . \tag{22}$$

Even though we now have the full data $\{N_1 \ldots N_n\}$, the likelihood depends only on the m quantities, $m < n$:

$$s_k \equiv \sum_{i=1}^{n} A_{ki} \frac{N_i}{N} \quad , \qquad i \leq k \leq m \tag{23}$$

which are therefore sufficient statistics. But these were just the given data $s_k = d_k$ in our original interpretation (13), and the maximum likelihood condition $\partial L/\partial \lambda_k = 0$ is again just Eq. (17), which now has three meanings!

Let us summarize these interesting connections. We started with no model and no sampling distribution, only prior information I_o which determined an hypothesis space consisting of an enumeration of the n^N conceivable outcomes of N trials, and some incomplete data $D = \{d_1 \ldots d_m\}$ that did not determine the frequencies $\{f_1 \ldots f_n\}$. Then recognition of the multiplicity factors led us to definite frequency estimates $\{\hat{f}_1 \ldots \hat{f}_n\}$, after all. But now, given this mathematical result (15) we may choose to ignore where it came from, and reinterpret it as a sampling distribution $p_i = p(x_i|\lambda_1 \ldots \lambda_m)$ with m parameters.

In this development, Eq. (17) has metamorphosed from the condition determining the Lagrange multipliers from the data in a variational problem, to a prediction of those data from the parameters in a sampling distribution, to the maximum-likelihood estimates of those parameters from different data. The writer finds it one of the most fascinating aspects of statistics--but also one of the greatest difficulties in teaching it--that totally different concepts and objectives may share the same mathematics.

In effect, then, the maximum-entropy principle has created a model and parameters for us, out of our prior information I_0 and incomplete data D. Every camp B maximum entropy problem defines what we shall call an adjunct model, usable in camp A.

But these adjunct models would be of little interest--one could hardly adjure others to use them--unless they had some desirable properties in their own right. Note first that a maximum-entropy distribution based on mean-value constraints is always in the exponential form (15) and so, by the Pitman-Koopman theorem the adjunct model is always one for which the generating data D would have been sufficient statistics. Pondering this may give one a deeper appreciation of the Shannon interpretation of the entropy expression $\Sigma p_i \log p_i$ as an "information measure".

We could give rather trivial examples, such as the analysis of Wolf's famous dice data, from this standpoint (Jaynes, 1978); but in view of space limitations let us proceed directly to a nontrivial case of current importance.

TIME SERIES

Nature generates a long time series $\{y_1 \ldots y_T\}$, but our data $D = \{d_0 \ldots d_m\}$ are incomplete, consisting only of the sample autocovariances up to a lag $m \ll T$. Given this information, what joint probability distribution $p(y_1 \ldots y_T)$ should we assign, and what estimate of the power spectrum should we make?

In camp A, the question does not seem to make sense. In camp B, with an hypothesis space on which to define our multiplicities and entropies, it becomes a well-posed problem, which will return an adjunct model to camp A.

In a real situation the values y_i are defined only to some finite accuracy $\pm\varepsilon$ over a finite range $|y| < Y$. Therefore the number n of possible realizations is finite, of the order of $(Y/\varepsilon)^T$. We shall take as our hypothesis space the n^N conceivable outcomes in N realizations of the time series.

The i'th realization $\{y_1^{(i)} \ldots y_T^{(i)}\}$ produces the sample autocovariance

$$R_k^{(i)} = \frac{1}{T} \sum_{t=1}^{T-k} y_t^{(i)} \, y_{t+k}^{(i)} \quad , \tag{24}$$

If we suppose for the moment that our data are the average values of this over N realizations, this fits into the combinatorial formalism (13)-(17) with A_{ki} proportional to (25), a convenient choice being

$$A_{ki} = \frac{T}{2} R_k^{(i)} \quad , \qquad \begin{matrix} 0 \leq k \leq m \\ 1 \leq i \leq n \end{matrix} \tag{25}$$

But ε is small and n, N large, so if we use the continuum approximation to the solution (15) we are committing no worse a sin than do those who use a continuous Chi-squared distribution, even though Chi-squared can take on only a discrete set of values.

Now the sum $\lambda_k A_{ki}$ in (19) is, from (25)

$$\sum_{k=0}^{m} \lambda_k A_{ki} = \frac{1}{2}(y' \Lambda y) \tag{26}$$

where $y = \{y_1 \ldots y_T\}$ a $(1 \times T)$ row vector, y' the $(T \times 1)$ column vector, and Λ the $(T \times T)$ matrix

$$\Lambda_{ij} = \left\{ \begin{matrix} \lambda_{|i-j|}, & -m \leq i,j \leq m \\ 0 \quad , & \text{otherwise} \end{matrix} \right\} \tag{27}$$

in the which λ_k are assembled in Toeplitz form. The adjunct distribution is therefore multivariate gaussian:

$$p(y_1 \cdots y_T | \lambda_0 \cdots \lambda_m) \propto \exp\left\{-\frac{1}{2}(y' \Lambda y)\right\} \quad . \tag{28}$$

An experienced entropy maximizer would proceed directly from the data to (26)-(28) without the long combinatorial argument we have appealed to--just as an experienced user of the Chi-squared test proceeds directly to the result without feeling the need to repeat Karl Pearson's original derivation of it every time it is used (but in both cases this facility in application may conceal the rationale of what is being done from someone not in on the secret).

Our adjunct model (28) is, except for an "end effect" factor, the sampling distribution of an autoregressive (AR) model of order m:

$$y_t + \sum_{k=1}^{m} a_k y_{t-k} = e_t \tag{29}$$

with $e_t \sim N(0,\sigma)$ and the AR coefficients related to the λ's by ($a_0 \equiv 1$):

$$\lambda_k = \sigma^{-2} \sum_{j=0}^{m-k} a_j \, a_{j+k} \tag{30}$$

Note that in this derivation we did not assume any "gaussian random process". The combinatorial argument told us that, of all distributions consistent with our data, the particular gaussian one (28) has the highest multiplicity--it could be realized by Nature in the greatest number of ways. Unless we had further information indicating a different distribution, then, it would seem irrational to use any other.

Also, we supposedfor the combinatorial argument that our data were averages over N realizations. Suppose we have data only from one realization--would we wish to use a different adjunct model to represent our knowledge of the process? One popular rationalization of our negative answer is that we can always imagine our one realization as cut up into N blocks, and the data are indeed averages over those blocks, so the situation is not fundamentally different. The point deserves further discussion, not given here; the adjunct model (28) can be obtained by completely different arguments based on Information Theory instead of combinatorics.

Of great current interest is the estimation of the power spectrum of a time series

$$S(\omega) \equiv \left| \sum_{t=1}^{T} y_t \, e^{-i\omega t} \right|^2 \quad . \tag{31}$$

The model (28) gives an estimate with an, at first glance, surprising form:

$$E[S(\omega)] = \frac{1}{\sum_{k=-m}^{m} \lambda_k \, e^{i\omega k}} \quad . \tag{32}$$

This is the basis of the Maximum Entropy Spectrum Analysis (MESA) method introduced by Burg (1967), which has largely supplanted older methods. As noted in our opening fable, MESA gives higher resolution without side-lobes because it takes into account prior information about multiplicity factors. But our derivation has supposed the data noiseless, and (32) needs to be modified when noise is present.

It was this spectrum analysis problem, more than anything else, that led to a clear perception of the camp A—camp B situation that this essay seeks to point out. The present status of MESA is presented in great detail in Haykin (1982).

This example illustrates several things about adjunct models. Put most briefly, adjunct models seem to anticipate--before we have seen the data--the models that sampling theorists eventually decide upon--after analyzing the data. The multivariate gaussian distribution (28) and/or the almost equivalent AR model (29) are just what people with experience in analyzing time series have been led to. Independently of Burg, Parzen (1968) and more recently many others, have advocated the same analytical form (32). In spite of their totally different language and philosophy, then, camp A and camp B do not necessarily differ in the final pragmatic results that they eventually arrive at.

Important truths can often be learned in more than one way; the MAXENT principle, a sufficiently deep intuition, or analysis of enough data could all lead us to (32). But intuition is unreliable, and data analysis is tedious, so MAXENT ought also to be in our bag of tools. Let us try to understand why this situation was inevitable.

RESOLUTION OF THE CONFLICT

In any experiment, certain factors are constant (under control) while others vary erratically, not under control. The resulting observable frequency distribution will be, almost certainly, the one that has maximum entropy subject to the constant factors--because it is a combinatorial theorem that the MAXENT distribution can be realized in far more ways than can any other obeying the same constraints.

With only a little poetic license, we could say that all real frequency distributions are MAXENT with respect to some constraints--and the most we could hope to learn from the experiment is: what are those fixed constraints?

This principle applies far beyond statistics--the "laws of physics" that we teach to our students with no mention of statistics, are actually no more than summaries of those aspects of physical phenomena that remain unchanged in varying situations.

From this standpoint it might appear that history--a record of unique events that will never be repeated--is the diametric opposite of physics. Yet for many the purpose of historical study is to detect those features that have been at work in all past civilizations, as a guide to the future. The "lessons of history" are only estimates of the constraints imposed on all civilizations by environment and human nature. And of course, the same can be said of psychology and econometrics. Behind our superficial differences there is a deep unity of logic and purpose. In saying this we are only extending what Karl Pearson pointed out long ago.

As we noted in the Introduction, when the sampling theorist chooses a model, he is expressing some kind of prior knowledge about the phenomenon. But the same model could express knowledge of a "mechanism"--or knowledge of multiplicity factors that, while not referring to any specific mechanism, tentatively suggest one.

If we have knowledge of--or if we hypothesize--some definite mechanism that tends to make a process repetitive by inducing correlations after a certain lag, then an autoregressive model would of course be the appropriate one to compare with our data, and to see this requires no entropy considerations [as we know from Volterra's ecological models, Richardson's "Statistics of Deadly Quarrels" or Yule's pendulum].

On the other hand, if we have no prior knowledge of the mechanism but we know that the available data will consist of correlations, then adjunct model considerations tell us that an autoregressive model is still appropriate, because it captures all the information about the process that is contained in the data. We see this as providing some support for those who, with such data at hand, have adopted an autoregressive model in a rather tentative and half-apologetic way; their choice has at least the justification that no other model could have made better use of the information they had.

But now we see that if different kinds of data later become available--other things than correlations--then a different adjunct model will be defined, with more parameters. At first glance it may smack of shifting sands to change our model when the type of data changes. But on second glance we see that the adjunct model is always the parsimonious one, introducing only the parameters that our data are able to deal with; and indeed, advancing to a new adjunct model only anticipates what the pure sampling theorist will do eventually; significance tests will undoubtedly show that the old model is not flexible enough to accommodate the new data, and so a new model with more parameters will be invented.

The point is that adjunct models tell us this in advance, and indeed in considerable detail. Equation (28) tells us not only the general form of the distribution but also the order of the corresponding AR model. It tells us something that seems obvious in retrospect but which, as recent literature shows, has not been obvious to all workers--sample autocovariance data only up to lag m can provide no evidence for the existence of AR coefficients beyond lag m. The Burg spectrum estimate (32) indicates this by the finite limits on the sum.

In contrast, the Blackman-Tukey (1958) estimate for the same data, implies non-zero AR coefficients far beyond the data--which were not indicated by the data and which were therefore, by our combinatorial arguments, confining us to an extremely small and unrepresentative subclass of all the spectra consistent with the data. The subclass happened to be one in which spurious "side-lobes" were present. Burg's removal of the spurious AR coefficients also removed the side-lobes, without losing any resolution. But the MAXENT approach would have told us this from the start.

Long ago, Harold Jeffreys (1939) and Jimmie Savage (1954, 1962) opined that Bayesian principles can often supply the missing theoretical justification for what sampling theorists do anyway, on intuitive grounds. Indeed, until an intuitive ad hockery has received some kind of theoretical justification we cannot judge its range of validity, or how to extend it.

But now we are suggesting more than that--prior information expressed by entropy factors gives us not just general theoretical justification, but quantitative details that the sampling theorist could discover eventually by analysis of enough data. Fitting a model to one's data so as to make the residuals as small as possible is the same thing as trying to account for as much of the data as possible by those constant influences that the model recognizes. If the model has captured all the systematic effects that are actually being used by Nature in generating the data, this fit will be judged in camp A as successful, reducing the residuals to a "purely random sequence" (although such a phrase is not in the vocabulary of camp B).

Note that the adjunct model is not asserted to be the "true" model. Indeed, that term has no meaning in camp B; the function of a model is only to represent our state of knowledge in the most honest way. The adjunct model is the one that yields the best predictions we are able to make from the prior information and the data D that generated it. In camp B it is a platitude that with different prior information, or a different kind of data, we have a different state of knowledge and therefore a different model.

Of course, if the data are of a wide enough scope to capture all the constraints that are being used by Nature, then the adjunct model will become "true" in the following sense, explained in more detail in Haykin (1982). If any new datum d_{m+1} is found to be only what we would have predicted from the old data, then $\lambda_{m+1} = 0$. If, after a certain point, all additional data are found to be thus redundant, then the adjuct model becomes stable, and further data do not affect our predictions. This is the reason why physicists are, as noted, lucky; our multiplicity factors are so overwhelming that our adjunct models become stable with very little data.

As in the case of multiple regression ($y = X\beta + e$), the number of algebraically independent parameters that can be estimated from our data is equal to the rank R of the matrix A (or X). The adjunct model recognizes this automatically. For example, if $R = m-2$, then it will be found that the λ's are connected by two algebraic relations, and two of them can be eliminated without changing the adjunct distribution. Here parsimony appears, not as an aesthetic consideration, but as a consequence of our honestly representing what we know -- and only what we know -- by avoiding gratituous assumptions like more AR coefficients than the data are able to "see". The Information Theory rationale for maximizing entropy (Jaynes, 1957) takes this "honesty" goal as the basic desideratum.

CONCLUSION

Finally, we return to that "more aggressive reply" promised earlier. It seems to be a common view that it is dangerous to express lack of definite prior knowledge by assigning equal prior probabilities. This seems to the writer to miss the whole point of scientific inference. In making an inference, we are not asserting that our prediction must be right; only that it is the best we can do on the information we have. Of course, it is possible that, unknown to us, Nature does have some counter-preferences that are not being properly taken into account if we use only multiplicity factors in our prior probabilities. It is also possible that our calculation of W was wrong because we made a bad guess about how Nature generates the scene.

But science does not advance on timidity. If our prior hypothesis is wrong or incomplete, how are we to discover that fact if we do not have the courage to go ahead and use it to make the best inferences we can; and compare them with observation? Indeed, it is only when our inferences turn out to be wrong that we have the opportunity to learn new things about Nature's workings.

This is not empty whistling in the dark; early in this century Gibbs made the best thermodynamic predictions possible on the information he had (multiplicity factors of classical mechanics); but at low temperatures Nature persisted in generating a scene with lower entropy than the one Gibbs predicted. Thus we learned that Nature has stronger constraints than are provided by classical theory; this was the first clue pointing the way to quantum theory. So getting a wrong answer is not such a calamity after all. But if Gibbs had lacked the courage to carry out his calculations for fear that he might be wrong, nobody would have realized that in low-temperature specific heats we had evidence for new laws of physics.

Today in astronomy, econometrics, geophysics, we have almost always some prior hypotheses about how our data are being generated by the real world. Suppose that, instead of "letting the data speak for themselves", we use this prior information in our inferences; what calamity is there to fear? If our hypotheses are right, we shall be rewarded by getting more reliable and accurate predictions, just as quantum statistics and maximum entropy image reconstruction and spectrum analysis are doing today.

If our hypotheses are significantly wrong, we shall obtain a far greater reward; new evidence about the true mechanism, that we would not otherwise have recognized. This rather favorable covenant that we have with Nature, inherent in the logic of scientific inference, could be exercised much more today if more workers recognized it.

The class of problems considered here is, of course, only a small fraction of all real statistical problems. Yet that class is wide enough to include many common and currently important problems. Prior information can never become a panacea, but it can sometimes make the process of inference more efficient. In cases where data acquisition is costly, in time or money, it is wasteful to have to learn from the data what we could have learned from neglected multiplicity factors. For a century, physicists and physical chemists have been able to make accurate and reliable thermodynamic predictions with very little data, simply because Josiah Willard Gibbs showed us how to include multiplicity in our calculations. Similar advantages are available to statisticians if they wish to use them.

This essay has sought to expound a viewpoint about the nature of inference that appears to the writer to combine the best features of both camp A and camp B, into something broader and more useful than either taken alone.

REFERENCES

J. Bertrand (1889); Calcul des probabilites, Gauthier-Villars, Paris

R. B. Blackman & J. W. Tukey (1958); The Measurement of Power Spectra, Dover Publishing Co., New York.

L. Boltzmann (1877); Wiener Berichte vol. 76, p. 373

J. P. Burg (1967); "Maximum Entropy Spectral Analysis", Proc. 37th Meeting, Society of Exploration Geophysicists: reprinted in Modern Spectrum Analysis, D. G. Childers, Editor, J. Wiley & Sons, Inc., New York (1978).

J. P. Burg (1975); Stanford University Doctoral Thesis

R. G. Currie (1981); "Solar Cycle Signal in Earth Rotation: Nonstationary Behavior", Science, vol. 211, pp 386-389

B. R. Frieden (1980); "Statistical Models for the Image Restoration Problem", Computer Graphics & Image Processing, Vol. 33, pp. 381-390

J. W. Gibbs (1875); "Heterogeneous Equilibrium", Conn. Acad. Sci.; reprinted, Dover Publications, Inc. (1961).

J. W. Gibbs (1902); Statistical Mechanics, Longmans, Green & Co.; reprinted, Dover Publications, Inc. (1961).

I. J. Good (1950);Probability and the Weighing of Evidence, Hafner Publishing Co., New York.

S. F. Gull & G. J. Daniell (1978); "Image reconstruction from incomplete and noisy data", Nature, Vol. 272, pp. 686-690

S. Haykin (ed.) (1982); IEEE Special Issue on Spectral Estimation. Contains many articles with further details.

E. T. Jaynes (1957); "Information Theory and Statistical Mechanics", Phys. Rev. vol 106, pp. 620-630; vol. 108, pp 171-190

E. T. Jaynes (1978); "Where do We Stand on Maximum Entropy?", in The Maximum-Entropy Formalism, R. D. Levine and M. Tribus, editors, M.I.T. Press, Cambridge MA; pp. 15-118

E. T. Jaynes (1980); "The Minimum Entropy Production Principle", in Annual Review of Physical Chemistry, Vol. 31, Annual Reviews, Inc., Palo Alto CA; pp. 579-601

H. Jeffreys (1939); Theory of Probability, Oxford University Press.

E. Parzen (1968); "Multiple Time Series Modelling", in Multivariate Analysis II Academic Press, New York.

H. Poincare (1912); Calcul des probabilites, Paris

L. J. Savage (1954); The Foundations of Statistics, J. Wiley & Sons, Inc., N. Y.

L. J. Savage (1962) discussion in The Foundations of Statistical Inference, M. S. Bartlett, Editor, Methuen & Co., Ltd., London