calgtut.tex, 7/8/1996

BAYESIAN METHODS: GENERAL BACKGROUND *

An Introductory Tutorial

E. T. Jaynes †
St. John's College and Cavendish Laboratory
Cambridge CB2 1TP, England

Abstract: We note the main points of history, as a framework on which to hang many background remarks concerning the nature and motivation of Bayesian/Maximum Entropy methods. Experience has shown that these are needed in order to understand recent work and problems. A more complete account of the history, with many more details and references, is given in Jaynes (1978).

The following discussion is essentially nontechnical; the aim is only to convey a little introductory "feel" for our outlook, purpose, and terminology, and to alert newcomers to common pitfalls of misunderstanding.

* Presented at the Fourth Annual Workshop on Bayesian/Maximum Entropy Methods, University of Calgary, August 1984. In the Proceedings Volume, *Maximum Entropy and Bayesian Methods in Applied Statistics*, J. H. Justice, Editor, Cambridge University Press (1985); pp. 1–25.

† Visiting Fellow, 1983–84. Permanent Address: Dept. of Physics, Campus Box 1105, Washington University, 1 Brookings Drive, St. Louis MO 63130, U.S.A.

HERODOTUS

The necessity of reasoning as best we can in situations where our information is incomplete is faced by all of us, every waking hour of our lives. We must decide what to do next, even though we cannot be certain what the consequences will be. Should I wear a raincoat today, eat that egg, cross that street, talk to that stranger, tote that bale, buy that book?

Long before studying mathematics we have all learned, necessarily, how to deal with such problems intuitively, by a kind of plausible reasoning where we lack the information needed to do the formal deductive reasoning of the logic textbooks. In the real world, some kind of extension of formal logic is needed.

And, at least at the intuitive level, we have become rather good at this extended logic, and rather systematic. Before deciding what to do, our intuition organizes the preliminary reasoning into stages: (I) Try to foresee all the possibilities that might arise; (II) Judge how likely each is, based on everything you can see and all your past experience; (III) In the light of this, judge what the probable consequences of various actions would be; (IV) Now make your decision.

From the earliest times this process of plausible reasoning preceding decisions has been recognized. Herodotus, in about 500 BC, discusses the policy decisions of the Persian kings. He notes that a decision was wise, even though it led to disastrous consequences, if the evidence at hand indicated it as the best one to make; and that a decision was foolish, even though it led to the happiest possible consequences, if it was unreasonable to expect those consequences.

So this kind of reasoning has been around for a long time, and has been well understood for a long time. Furthermore, it is so well organized in our minds in qualitative form that it seems obvious that: (A) the above stages of reasoning can be reproduced in quantitative form by a mathematical model; (B) such an extended logic would be very useful in such areas as science, engineering, and economics, where we are also obliged constantly to reason as best we can in spite of incomplete information, but the number of possibilities and amount of data are far too great for intuition to keep track of.

BERNOULLI

A serious, and to this day still useful, attempt at a mathematical representation was made by James Bernoulli (1713), who called his work *Ars Conjectandi*, or "The Art of Conjecture", a name that might well be revived today because it expresses so honestly and accurately what we are doing. But even though it is only conjecture, there are still wise and foolish ways, consistent and inconsistent ways, of doing it. Our extended logic should be, in a sense that was made precise only much later, an optimal or "educated" system of conjecture.

First one must invent some mathematical way of expressing a state of incomplete knowledge, or information. Bernoulli did this by enumerating a set of basic "equally possible" cases which we may denote by $\{x_1, x_2, \ldots, x_N\}$, and which we may call, loosely, either events or propositions. This defines our "field of discourse" or "hypothesis space" H_0. If we are concerned with two tosses of a die, $N = 6^2 = 36$.

Then one introduces some proposition of interest A, defined as being true on some specified subset $H(A)$ of M points of H_0, false on the others. M, the "number of ways" in which A could be true, is called the *multiplicity* of A, and the *probability* of A is defined as the proportion

$$p(A) = M/N \ .$$

The rules of reasoning consist of finding the probabilities $p(A)$, $p(B)$, *etc.* of different propositions by counting the number of ways they can be true. For example, the probability that both A and B are true is the proportion of H_0 on which both are true. More interesting, if we learn that A is true, our hypothesis space contracts to $H(A)$ and the probability of B is changed to the proportion of $H(A)$ on which B is true. If we then learn that B is false, our hypothesis space may contract further, changing the probability of some other proposition C, and so on.

Such elementary rules have an obvious correspondence with common sense, and they are powerful enough to be applied usefully, not only in the game of "Twenty Questions", but in some quite substantial problems of reasoning, requiring nontrivial combinatorial calculations. But as Bernoulli recognized, they do not seem applicable to all problems; for while we may feel that we know the appropriate H_0 for dice tossing, in other problems we often fail to see how to define any set H_0 of elementary "equally possible" cases. As Bernoulli put it, "What mortal will ever determine the number of diseases?" How then could we ever calculate the probability of a disease?

Let us deliver a short Sermon on this. Faced with this problem, there are two different attitudes one can take. The conventional one, for many years, has been to give up instantly and abandon the entire theory. Over and over again modern writers on statistics have noted that no general rule for determining the correct H_0 is given, *ergo* the theory is inapplicable and into the waste–basket it goes.

But that seems to us a self–defeating attitude that loses nearly all the value of probability theory by missing the point of the game. After all, our goal is not omniscience, but only to reason as best we can with whatever incomplete information we have. To demand more than this is to demand the impossible; neither Bernoulli's procedure nor any other that might be put in its place can get something for nothing.

The reason for setting up H_0 is not to describe the Ultimate Realities of the Universe; that is unknown and remains so. By definition, the function of H_0 is to represent what we know; it cannot be unknown. So a second attitude recommends itself; define your H_0 as best you can – all the diseases you know – and get on with the calculations.

Usually this suggestion evokes howls of protest from those with conventional training; such calculations have no valid basis at all, and can lead to grotesquely wrong predictions. To trust the results could lead to calamity.

But such protests also miss the point of the game; they are like the reasoning of a chess player who thinks ahead only one move and refuses to play at all unless the next move has a guaranteed win. If we think ahead two moves, we can see the true function and value of probability theory in inference.

When we first define H_0, because of our incomplete information we cannot be sure that it really expresses all the possibilities in the real world. Nor can we be sure that there is no unknown symmetry–breaking influence at work making it harder to realize some possibilities than others. If we knew of such an influence, then we would not consider the x_i equally likely.

To put it in somewhat anthropomorphic terms, we cannot be sure that our hypothesis space H_0 is the same as Nature's hypothesis space H_N. The conventional attitude holds that our calculation is invalid unless we know the "true" H_N; but that is something that we shall never know. So, ignoring all protests, we choose to go ahead with that shaky calculation from H_0, which is the best we can actually do. What are the possible results?

Suppose our predictions turn out to be right; *i.e.* out of a certain specified set of propositions $(A_1, A_2, \ldots, A_m)$, the one A_k that we thought highly likely to be true (because it is true on a much larger subset of H_0 than any other) is indeed confirmed by observation. That does not prove that our H_0 represented correctly all those, and only those, possibilities that exist in Nature, or that no symmetry–breaking influences exist. But it does show that our H_0 is not sufficiently different from Nature's H_N to affect this prediction. Result: the theory has served a useful predictive purpose, and we have more confidence in our H_0. If this success continues with many different sets of propositions, we shall end up with very great confidence in H_0. Whether it is "true" or not, it has predictive value.

But suppose our prediction does indeed turn out to be grotesquely wrong; Nature persists in generating an entirely different A_j than the one we favored. Then we know that Nature's H_N is importantly different from our H_0, and the nature of the error gives us a clue as to how they differ. As a result, we are in a position to define a better hypothesis space H_1, repeat the calculations to see what predictions it makes, compare them with observation, define a still better H_2, and so on.

Far from being a calamity, this is the essence of the scientific method. H_0 is only our unavoidable starting point.

As soon as we look at the nature of inference at this many–moves–ahead level of perception, our attitude toward probability theory and the proper way to use it in science becomes almost diametrically opposite to that expounded in most current textbooks. We need have no fear of making shaky calculations on inadequate knowledge; for if our predictions are indeed wrong, then we shall have an opportunity to improve that knowledge, an opportunity that would have been lost had we been too timid to make the calculations.

Instead of fearing wrong predictions, we look eagerly for them; it is only when predictions based on our present knowledge fail that probability theory leads us to fundamental new knowledge.

Bernoulli implemented the point just made, and in a more sophisticated way than we supposed in that little Sermon. Perceiving as noted that except in certain gambling devices like dice we almost never know Nature's set H_N of possibilities, he conceived a way of probing H_N, in the case that one can make repeated independent observations of some event A; for example, administering a medicine to many sick patients and noting how many are cured.

We all feel intuitively that, under these conditions, events of higher probability M/N should occur more often. Stated more carefully, events that have higher probability on H_0 should be predicted to occur more often; events with higher probability on H_N should be observed to occur more often. But we would like to see this intuition supported by a theorem.

Bernoulli proved the first mathematical connection between probability and frequency, today known as the weak law of large numbers. If we make n independent observations and find A true m times, the observed frequency $f(A) = m/n$ is to be compared with the probability $p(A) = M/N$. He showed that in the limit of large n, it becomes practically certain that $f(A)$ is close to $p(A)$. Laplace showed later that as n tends to infinity the probability remains more than 1/2 that $f(A)$ is in the shrinking interval $p(A) \pm q$, where $q^2 = p(1-p)/n$.

There are some important technical qualifications to this, centering on what we mean by "independent"; but for present purposes we note only that often an observed frequency $f(A)$ is in some sense a reasonable estimate of the ratio M/N in Nature's hypothesis space H_N. Thus we have, in many cases, a simple way to test and improve our H_0 in a semiquantitative way. This was an important start; but Bernoulli died before carrying the argument further.

BAYES

Thomas Bayes was a British clergyman and amateur mathematician (a very good one – it appears that he was the first person to understand the nature of asymptotic expansions), who died in 1761. Among his papers was found a curious unpublished manuscript. We do not know what he intended to do with it, or how much editing it then received at the hands of others; but it was published in 1763 and gave rise to the name "Bayesian Statistics". For a photographic reproduction of the work as published, with some penetrating historical comments, see Molina (1963). It gives, by lengthy arguments that are almost incomprehensible today, a completely different kind of solution to Bernoulli's unfinished problem.

Where Bernoulli had calculated the probability, given N, n, and M, that we would observe A true m times (what is called today the "sampling distribution"), Bayes turned it around and gave in effect a formula for the probability, given N, n, and m, that M has various values. The method was long called "inverse probability". But Bayes' work had little if any direct influence on the later development of probability theory.

LAPLACE

In almost his first published work (1774), Laplace rediscovered Bayes' principle in greater clarity and generality, and then for the next 40 years proceeded to apply it to problems of astronomy, geodesy, meteorology, population statistics, and even jurisprudence. The basic theorem appears today as almost trivially simple; yet it is by far the most important principle underlying scientific inference.

Denoting various propositions by A, B, C, *etc.*, let AB stand for the proposition "both A and B are true", $\overline{A}$ = "A is false", and let the symbol $p(A|B)$ stand for "the probability that A is true, given that B is true". Then the basic product and sum rules of probability theory, dating back in essence to before Bernoulli, are

$$p(AB|C) = p(A|BC)\,p(B|C) \tag{1}$$

$$p(A|B) + p(\overline{A}|B) = 1 \tag{2}$$

But AB and BA are the same proposition, so consistency requires that we may interchange A and B in the right–hand side of (1). If $p(B|C) > 0$, we thus have what is always called "Bayes' Theorem" today, although Bayes never wrote it:

$$p(A|BC) = p(A|C)\,\frac{p(B|AC)}{p(B|C)} \tag{3}$$

But this is nothing more than the statement that the product rule is consistent; why is such a seeming triviality important?

In (3) we have a mathematical representation of the process of learning; exactly what we need for our extended logic. $p(A|C)$ is our "prior probability" for A, when we know only C. $p(A|BC)$ is its "posterior probability", updated as a result of acquiring new information B. Typically, A represents some hypothesis, or theory, whose truth we wish to ascertain, B represents new data from some observation, and the "prior information" C represents the totality of what we knew about A before getting the data B.

For example – a famous example that Laplace actually did solve – proposition A might be the statement that the unknown mass M_S of Saturn lies in a specified interval, B the data from observatories about the mutual perturbations of Jupiter and Saturn, C the common sense observation that M_S cannot be so small that Saturn would lose its rings; or so large that Saturn would disrupt the solar system. Laplace reported that, from the data available up to the end of the 18'th Century, Bayes' theorem estimates M_S to be (1/3512) of the solar mass, and gives a probability of .99991, or odds of 11,000:1, that M_S lies within 1% of that value. Another 150 years' accumulation of data has raised the estimate 0.63 percent.

The more we study it, the more we appreciate how nicely Bayes' theorem corresponds to, and improves on, our common sense. In the first place, it is clear that the prior probability $p(A|C)$ is necessarily present in all inference; to ask "What do you know about A after seeing the data B?" cannot have any definite answer – because it is not a well–posed question – if we fail to take into account, "What did you know about A before seeing B?"

Even this platitude has not always been perceived by those who do not use Bayes' theorem and go under the banner: "Let the data speak for themselves!" They cannot, and never have. If we want to decide between various possible theories but refuse to supplement the data with prior information about them, probability theory will lead us inexorably to favor the "Sure Thing" theory ST, according to which every minute detail of the data was inevitable; nothing else could possibly have happened. For the data always have a much higher probability [namely $p(D|ST) = 1$] on ST than on any other theory; ST is always the orthodox "maximum likelihood" solution over the class of all theories. Only our extremely low prior probability for ST can justify rejecting it.

Secondly, we can apply Bayes' theorem repeatedly as new pieces of information $B_1, B_2, \cdots$ are received from the observatories, the posterior probability from each application becoming the prior probability for the next. It is easy to verify that (3) has the chain consistency that common sense would demand; at any stage the probability that Bayes' theorem assigns to A depends only on the total evidence $B_{tot} = B_1 B_2 \cdots B_k$ then at hand, not on the order in which the different updatings happened. We could reach the same conclusion by a single application of Bayes' theorem using B_{tot}.

But Bayes' theorem tells us far more than intuition can. Intuition is rather good at judging what pieces of information are relevant to a question, but very unreliable in judging the relative cogency of different pieces of information. Bayes' theorem tells us quantitatively just how cogent every piece of information is.

Bayes' theorem is such a powerful tool in this extended logic that, after 35 years of using it almost daily, I still feel a sense of excitement whenever I start on a new, nontrivial problem; because I know that before the calculation has reached a dozen lines it will give me some important new insight into the problem, that nobody's intuition has seen before. But then that surprising result always seems intuitively obvious after a little meditation; if our raw intuition was powerful enough we would not need extended logic to help us.

Two examples of the fun I have had doing this, with full technical details, are in my papers "Bayesian Spectrum and Chirp Analysis" given at the August 1983 Laramie Workshop, and "Highly Informative Priors" to appear in the Proceedings Volume for the September 1983 International Meeting on Bayesian Statistics, Valencia, Spain. In both cases, completely unexpected new insight from Bayes' theorem led to quite different new methods of data analysis and more accurate results, in two problems (spectrum analysis and seasonal adjustment) that had been treated for decades by non–Bayesian methods. The Bayesian analysis had the technical means to take into account some previously neglected prior information.

Laplace, equally aware of the power of Bayes' theorem, used it to help him decide which astronomical problems to work on. That is, in which problems is the discrepancy between prediction and observation large enough to give a high probability that there is something new to be found? Because he did not waste time on unpromising research, he was able in one lifetime to make more of the important discoveries in celestial mechanics than anyone else.

Laplace also published (1812) a remarkable two–volume treatise on probability theory in which the analytical techniques for Bayesian calculations were developed to a level that is seldom surpassed today. The first volume contains, in his methods for solving finite difference equations, almost all of the mathematics that we find today in the theory of digital filters.

Yet all of Laplace's impressive accomplishments were not enough to establish Bayesian analysis in the permanent place that it deserved in science. For more than a Century after Laplace, we were deprived of this needed tool by what must be the most disastrous error of judgment ever made in science.

In the end, all of Laplace's beautiful analytical work and important results went for naught because he did not explain some difficult conceptual points clearly enough. Those who came after him got hung up on inability to comprehend his rationale and rejected everything he did, even as his masses of successful results were staring them in the face.

JEFFREYS

Early in this Century, Sir Harold Jeffreys rediscovered Laplace's rationale and, in the 1930's, explained it much more clearly than Laplace did. But still it was not comprehended; and for thirty more years Jeffreys' work was under attack from the very persons who had the most to gain by understanding it (some of whom were living and eating with him daily here in St. John's College, and had the best possible opportunity to learn from him). But since about 1960 comprehension of what Laplace and Jeffreys were trying to say has been growing, at first slowly and today quite rapidly.

This strange history is only one of the reasons why, today, we Bayesians need to take the greatest pains to explain our rationale, as I am trying to do here. It is not that it is technically complicated; it is the way we have all been thinking intuitively from childhood. It is just so different from what we were all taught in formal courses on "orthodox" probability theory, which paralyze the mind into an inability to see the distinction between probability and frequency. Students who come to us free of that impediment have no difficulty in understanding our rationale, and are incredulous that anyone could fail to comprehend it.

My Sermons are an attempt to spread the message to those who labor under this handicap. Summarizing Bernoulli's work was the excuse for delivering the first Sermon establishing, so to speak, our Constitutional Right to use H_0 even if it may not be the same as H_N. Now Laplace and Jeffreys inspire our second Sermon, on how to choose H_0 given our prior knowledge; a matter on which they made the essential start. To guide us in this choice there is a rather fundamental "Desideratum of Consistency": *In two problems where we have the same state of knowledge, we should assign the same probabilities.*

As an application of this desideratum, if the hypothesis space H_0 has been chosen so that we have no information about the x_i beyond their enumeration, then as an elementary matter of symmetry the only consistent thing we can do is to assign equal probability to all of them; if we did anything else, then by a mere permutation of the labels we could exhibit a second problem in which our state of knowledge is the same, but in which we are assigning different probabilities.

This rationale is the first example of the general group invariance principle for assigning prior probabilities to represent "ignorance". Although Laplace used it repeatedly and demonstrated its successful consequences, he failed to explain that it is not arbitrary, but required by logical consistency to represent a state of knowledge. Today, 170 years later, this is still a logical pitfall that causes conceptual hangups and inhibits applications of probability theory.

Let us emphasize that we are using the word "probability" in its original – therefore by the usual scholarly standards correct – meaning, as referring to incomplete human information. It has, fundamentally, nothing to do with such notions as "random variables" or "frequencies in random experiments"; even the notion of "repetition" is not necessarily in our hypothesis space.

In cases where frequencies happen to be relevant to our problem, whatever connections they may have with probabilities appear automatically, as mathematically derived consequences of our extended logic (Bernoulli's limit theorem being the first example). But, as shown in a discussion of fluctuations in time series (Jaynes, 1978), those connections are often of a very different nature than is supposed in conventional pedagogy; the predicted mean–square fluctuation is not the same as the variance of the first–order probability distribution.

So to assign equal probabilities to two events is not in any way an assertion that they must occur equally often in any "random experiment"; as Jeffreys emphasized, it is only a formal way of saying "I don't know". Events are not necessarily capable of repetition; the event that the mass of Saturn is less than (1/3512) had, in the light of Laplace's information, the same probability as the event that it is greater than (1/3512), but there is no "random experiment" in which we expect those events to occur equally often. Of course, if our hypothesis space is large enough to accommodate the repetitions, we can calculate the *probability* that two events occur equally often.

To belabor the point, because experience shows that it is necessary: In our scholastically correct terminology, a *probability* p is an abstract concept, a quantity that we *assign* theoretically, for the purpose of representing a state of knowledge, or that we *calculate* from previously assigned probabilities using the rules (1) – (3) of probability theory. A *frequency* f is, in situations where it makes sense to speak of repetitions, a factual property of the real world, that we *measure* or *estimate*. So instead of committing the error of saying that the probability *is* the frequency, we ought to calculate the probability $p(f)\,df$ that the frequency lies in various intervals df – just as Bernoulli did.

In some cases our information, although incomplete, still leads to a very sharply peaked probability distribution $p(f)$; and then we can indeed make very confident predictions of frequencies. In these cases, if we are not making use of any information other than frequencies, our conclusions will agree with those of "random variable" probability theory as usually taught today. Our results do not conflict with frequentist results whenever the latter are justified. From a pragmatic standpoint (*i.e.*, ignoring philosophical stances and looking only at the actual results), "random variable" probability theory is contained in the Laplace–Jeffreys theory as a special case.

But the approach being expounded here applies also to many important real problems – such as the "pure generalized inverse" problems of concern to us at this Workshop – in which there is not only no "random experiment" involved, but we have highly cogent information that must be taken into account in our probabilities, but does not consist of frequencies.

A theory of probability that fails to distinguish between the notions of probability and frequency is helpless to deal with such problems. This is the reason for the present rapid growth of Bayesian methods – which can deal with them, and with demonstrated success. And of course, we can deal equally well with the compound case where we have both random error and cogent non–frequency information.

COX

One reason for these past problems is that neither Laplace nor Jeffreys gave absolutely compelling arguments – that would convince a person who did not want to believe it – proving that the rules (1) – (3) of probability theory are uniquely favored, by any clearly stated criterion of optimality, as the "right" rules for conducting inference. To many they appeared arbitrary, no better than a hundred other rules one could invent. But those rules were – obviously and trivially – valid rules for combining frequencies, so in the 19'th Century the view arose that a probability is not respectable unless it is also a frequency.

In the 1930's the appearance of Jeffreys' work launched acrimonious debates on this issue. The frequentists took not the slightest note of the masses of evidence given by Laplace and Jeffreys, demonstrating the pragmatic success of those rules when applied without the sanction of any frequency connection; they had their greatest success in just the circumstances where the frequentists held them to be invalid.

Into this situation there came, in 1946, a modest little paper by Richard T. Cox, which finally looked at the problem in just the way everybody else should have. He issued no imperatives declaring that rules (1) – (3) were or were not valid for conducting inference. Instead he observed that, whether or not Laplace gave us the right "calculus of inductive reasoning", we can at least raise the question whether such a calculus could be created today.

Supposing that degrees of plausibility are to be represented by real numbers, he found the conditions that such a calculus be consistent (in the sense that if two different methods of calculation are permitted by the rules, then they should yield the same result). These consistency conditions took the form of two functional equations, whose general solutions he found. Those solutions uniquely determined the rules (1) and (2), to within a change of variables that can alter their form but not their content.

So, thanks to Cox, it was now a theorem that any set of rules for conducting inference, in which we represent degrees of plausibility by real numbers, is necessarily either equivalent to the Laplace–Jeffreys rules, or inconsistent. The reason for their pragmatic success is then pretty clear. Those who continued to oppose Bayesian methods after 1946 have been obliged to ignore not only the pragmatic success, but also the theorem.

SHANNON

Two years later, Claude Shannon (1948) used Cox's method again. He sought a measure of the "amount of uncertainty" in a probability distribution. Again the conditions of consistency took the form of functional equations, whose general solution he found. The resulting measure proved to be

$$S_I = -\sum p_i \log p_i \,,$$

just what physicists had long since related to the entropy of thermodynamics.

Gibbs (1875) had given a variational principle in which maximization of the phenomenological "Clausius entropy" S_E led to all the useful predictions of equilibrium thermodynamics. But S_E still had to be determined by calorimetric measurements (the familiar $S_E = \int dQ/T$ over a reversible path that science students learn about). Boltzmann (1877), Gibbs (1902) and von Neumann (1928) gave three variational principles in which the maximization of the "Shannon information entropy" S_I led, in both classical and quantum theory, to a theoretical prediction of the Clausius entropy S_E; and thus again (if one was a good enough calculator) to all the useful results of equilibrium thermodynamics – without any need for the calorimetric measurements.

But again we had been in a puzzling situation just like that of Bayesian inference. Also here we had (from Gibbs) a simple, mathematically elegant formalism that led, in the quantum theory version, to enormous pragmatic success; but had no clearly demonstrated theoretical justification. Also here, arguments over its rationale – among others, does it or does it not require "ergodic" properties of the equations of motion? – had been underway for decades, the partisans on each side doubting the sanity of those on the other. But again, Shannon's consistency theorem finally showed in what sense entropy maximization generated optimal inferences. Whether or not the system is ergodic, the formalism still yields the best predictions that could have been made from the information we had.

This was really a repetition of history from Bernoulli. Shannon's theorem only established again, in a new area, our Constitutional Right to use H_0 based on whatever information we have, whether or not it is the same as H_N. Our previous remarks about many–moves–ahead perception apply equally well here; if our H_0 differs from H_N, how else can we discover that fact but by having the courage to go ahead with the calculations on H_0 to see what predictions it makes?

Gibbs had that courage; his H_0 of classical mechanics was "terrible" by conventional attitudes, for it predicted only equations of state correctly, and gave wrong specific heats, vapor pressures, and equilibrium constants. Those terribly wrong predictions were the first clues pointing to quantum theory; our many–moves–ahead scenario is not a fancy, but an historical fact. It is clear from Gibbs' work of 1875 that he understood that scenario; but like Laplace he did not explain it clearly, and it took a long time for it to be rediscovered.

There is one major difference in the two cases. The full scope and generality of Bayesian inference had been recognized already by Jeffreys, and Cox's theorem only legitimized what he had been doing all along. But the new rationale from Shannon's theorem created an enormous expansion of the scope of maximum entropy.

It was suddenly clear that, instead of applying only to prediction of equilibrium thermodynamics, as physicists had supposed before Shannon, the variational principles of Gibbs and von Neumann extended as well to nonequilibrium thermodynamics, and to any number of new applications outside of thermodynamics. As attested by the existence of this Workshop, it can be used in spectrum analysis, image reconstruction, crystallographic structure determination, econometrics; and indeed any problem, whatever the subject–matter area, with the following logical structure:

We can define an hypothesis space H_0 by enumerating some perceived possibilities $(x_1 \cdots x_N)$; but we do not regard them as equally likely, because we have some additional evidence E. It is not usable as the "data" B in Bayes' theorem (3) because E is not an event in H_0 and does not have a "sampling distribution" $p(E|C)$. But E leads us to impose some constraint on the probabilities $p_i = p(x_i)$ that we assign to the elements of H_0, which forces them to be nonuniform, but does not fully determine them (the number of constraints is less than N).

We interpret Shannon's theorem as indicating that, out of all distributions p_i that agree with the constraints, the one that maximizes the Shannon entropy represents the "most honest" description of our state of knowledge, in the following sense: it expresses the enumeration of the possibilities and the evidence E; but is careful to assume nothing beyond that.

If we subsequently acquire more information B that can be interpreted as an event in H_0, then we can update this distribution by Bayes' theorem. In other words, MAXENT has given us the means to escape from the "equally possible" domain of Bernoulli and Laplace, and construct nonuniform prior probability distributions.

Thus came the unification of these seemingly different fields. Boltzmann and Gibbs had been, unwittingly, solving the prior probability problem of Bernoulli and Laplace, in a very wide class of problems. The area of useful applications this opens up will require 100 years to explore and exploit.

But this has only scratched the surface of what can be done in inference, now that we have escaped from at least some past errors. We can see, but only vaguely, still more unified, more powerful, and more general theories of inference which will regard our present one as an approximate special case.

COMMUNICATION DIFFICULTIES

Our background remarks would be incomplete without taking note of a serious disease that has afflicted probability theory for 200 years. There is a long history of confusion and controversy, leading in some cases to a paralytic inability to communicate. This has been caused, not only by confusion over the notions of probability and frequency, but even more by repeated failure to distinguish between different problems that happen to lead to similar mathematics. We are concerned here with only one of these failures.

Starting with the debates of the 1930's between Jeffreys and Fisher in the British Statistical Journals, there has been a puzzling communication block that has prevented orthodoxians from comprehending Bayesian methods, and Bayesians from comprehending orthodox criticisms of our methods. On the topic of how probability theory should be used in inference, L. J. Savage (1954) remarked that "*there has seldom been such complete disagreement and breakdown of communication since the Tower of Babel*".

For example, the writer recalls reading, as a student in the late 1940's, the orthodox textbook by H. Cramér (1946). His criticisms of Jeffreys' approach seemed to me gibberish, quite unrelated to what Jeffreys had done. There was no way to reply to the criticisms, because they were just not addressed to the topic. The nature of this communication block has been realized only quite recently.

For decades Bayesians have been accused of "supposing that an unknown parameter is a random variable"; and we have denied hundreds of times, with increasing vehemence, that we are making any such assumption. We have been unable to comprehend why our denials have no effect, and that charge continues to be made.

Sometimes, in our perplexity, it has seemed to us that there are two basically different kinds of mentality in statistics; those who see the point of Bayesian inference at once, and need no explanation; and those who never see it, however much explanation is given.

But a Seminar talk by Professor George Barnard, given in Cambridge in February 1984, provided a clue to what has been causing this Tower of Babel situation. Instead of merely repeating the old accusation (that we could only deny still another time), he expressed the orthodox puzzlement over Bayesian methods in a different way, more clearly and specifically than we had ever heard it put before.

Barnard complained that Bayesian methods of parameter estimation, which present our conclusions in the form of a posterior distribution, are illogical; for "*How could the distribution of a parameter possibly become known from data which were taken with only one value of the parameter actually present?*"

This extremely revealing comment finally gave some insight into what has been causing our communication problems. Bayesians have always known that orthodox terminology is not well adapted to expressing Bayesian ideas; but at least this writer had not realized how bad the situation was.

Orthodoxians trying to understand Bayesian methods have been caught in a semantic trap by their habitual use of the phrase "distribution of the parameter" when one should have said "distribution of the probability". Bayesians had supposed this to be merely a figure of speech; *i.e.* that those who used it did so only out of force of habit, and really knew better. But now it seems that our critics have been taking that phraseology quite literally all the time.

Therefore, let us belabor still another time what we had previously thought too obvious to mention. In Bayesian parameter estimation, both the prior and posterior distributions represent, not any measurable property of the parameter, but only our own state of knowledge about it. The width of the distribution is not intended to indicate the range of variability of the true values of the parameter, as Barnard's terminology led him to suppose. It indicates the range of values that are consistent with our prior information and data, and which honesty therefore compels us to admit as possible values. What is "distributed" is not the parameter, but the probability.

Now it appears that, for all these years, those who have seemed immune to all Bayesian explanation have just misunderstood our purpose. All this time, we had thought it clear from our subject–matter context that we are trying to estimate the value that the parameter had *when the data were taken.* Put more generally, we are trying to draw inferences about what actually did happen in the experiment; not about the things that might have happened but did not.

But it seems that, all this time, our critics have been trying to interpret our calculations in a different way, imposed on them by their habits of terminology, as an attempt to solve an entirely different problem. With this realization, our past communication difficulties become understandable: the problem our critics impute to us has – as they correctly see – no solution from the information at hand. The fact that we nevertheless get a solution then seems miraculous to them, and we are accused of trying to get something for nothing.

Re–reading Cramér and other old debates in the literature with this in mind, we can now see that this misunderstanding of our purpose was always there, but covertly. Our procedure is attacked on the overt grounds that our prior probabilities are not frequencies – which seemed to

12

us a mere philosophical stance, and one that we rejected. But to our critics this was far more than a philosophical difference; they took it to mean that, lacking the information to solve the problem, we are contriving a false solution.

What was never openly brought out in these frustrating debates was that our critics had in mind a completely different problem than the one we were solving. So to an orthodoxian our denials seemed dishonest; while to a Bayesian, the orthodox criticisms seemed so utterly irrelevant that we could see no way to answer them. Our minds were just operating in different worlds.

Perhaps, realizing this, we can now see a ray of light that might, with reasonable good will on both sides, lead to a resolution of our differences. In the future, it will be essential for clear communication that all parties see clearly the distinction between these two problems:

(I) In the Bayesian scenario we are *estimating*, from our prior information and data, the unknown constant value that the parameter had when the data were taken.

(II) In Barnard's we are *deducing*, from prior knowledge of the frequency distribution of the parameter over some large class C of repetitions of the whole experiment, the frequency distribution that it has in the subclass $C(D)$ of cases that yield the same data D.

The problems are so different that one might expect them to be solved by different procedures. Indeed, if they had led to completely different algorithms, the problems would never have been confused and we might have been spared all these years of controversy. But it turns out that both problems lead, via totally different lines of reasoning, to the same actual algorithm; application of Bayes' theorem.

However, we do not know of any case in which Bayes' theorem has actually been used for problem (II); nor do we expect to hear of such a case, for three reasons: (a) In real problems, the parameter of interest is almost always an unknown constant, not a "random variable"; (b) Even if it is a random variable, what is of interest is almost always the value it had during the real experiment that was actually done; not its frequency distribution in an imaginary subclass of experiments that were not done; (c) Even if that imaginary frequency distribution were the thing of interest, we never know the prior frequency distribution that would be needed.

That is, even if the orthodoxian wanted to solve problem (II), he would not have, any more than we do, the information needed to use Bayes' theorem for it, and would be obliged to seek some other method.

This unfortunate mathematical accident reinforced the terminological confusion; when orthodoxians saw us using the Bayesian algorithm, they naturally supposed that we were trying to solve problem (II). Still more reinforcement came from the fact that, since orthodoxy sees no meaning in a probability which is not also a frequency, it is unable to see Bayes' theorem as the proper procedure for solving problem (I). Indeed for that problem it is obliged to seek, not just other methods than Bayes' theorem; but other tools than probability theory. Those are the intuitive *ad hockeries* that orthodox statistics abounds with.

For the Bayesian, who does see meaning in a probability that is not a frequency, all the theoretical principles needed to solve problem (I) are contained in the product and sum rules (1), (2) of probability theory. He views them, not merely as rules for calculating frequencies (which they are, but trivially); but also rules for conducting inference – a nontrivial property requiring mathematical demonstration.

But orthodoxians do not read the Bayesian literature, in which it is a long since demonstrated fact, not only in the work of Cox but in quite different approaches by de Finetti, Wald, Savage, Lindley, and others, that Bayesian methods yield the optimal solution to problem (I), by some very basic, and it seems to us inescapable, criteria of optimality. Pragmatically, our numerical results confirm this; whenever the Bayesian and orthodoxian arrive at different results for problem (I), closer examination has always shown the Bayesian result to be superior (Jaynes, 1976).

In view of this, we are not surprised to find that, while orthodox criticisms of Bayesian methods deplore our philosophy and procedure, they always stop short of examining our actual numerical results in real problems and comparing them with orthodox results (when the latter exist). The few who take the trouble to do this quickly become Bayesians themselves.

Today, it is our pragmatic results, far more than the philosophy or even the optimality theorems, that is making Bayesianity a rapidly growing concern, taking over one after another of the various areas of scientific inference. The era of the computer has brought out the facts of actual performance, in a way that can no longer be obscured by arguments over philosophy.

So, rising above past criticisms which now appear to have been only misunderstandings of our purpose, Bayesians are in a position to help orthodox statistics in its current serious difficulties. In the real problems of scientific inference, it is almost invariably problem (I) that is in need of solution. But aside from a few special (and to the Bayesian, trivial) cases, orthodoxy has no satisfactory way to deal with problem (I).

In our view, then, the orthodoxian has a great deal to gain in useful results (and as far as we can see, nothing to lose but his ideological chains) by joining the Bayesian camp and finally taking advantage of this powerful tool. Failing to do this, he faces obsolescence as new applications of the kind we are discussing at this Workshop pass far beyond his domain.

There have been even more astonishing misunderstandings and misrepresentations of the nature and purpose of the Maximum Entropy principle. These troubles just show that statistics is different from other fields. Our problems are not only between parties trying to communicate with each other; not all readers are even trying to understand your message. It's a jungle out there, full of predators tensed like coiled springs, ready and eager to pounce upon every opportunity to misrepresent your meaning, your purpose, and even your results. I am by no means the first to observe this; both H. Jeffreys and R. A. Fisher complained about it in the 1930's.

These are the main points of the historical background of this field; but the important background is not all historical. The general nature of Bayesian/Maximum Entropy methods still needs some background clarification, because their logical structure is different not only from that of orthodox statistics, but also from that of conventional mathematical theories in general.

IS OUR LOGIC OPEN OR CLOSED?

Let us dwell a moment on a circumstance that comes up constantly in this field. It cannot be an accident that almost all of the literature on Bayesian methods of inference is by authors concerned with specific real problems of the kind that arise in science, engineering, and economics; while most mathematicians concerned with probability theory take no note of the existence of these methods. Mathematicians and those who have learned probability theory from them seem uncomfortable with Bayesianity, while physicists take to it naturally. This might be less troublesome if we understood the reason for it.

Perhaps the answer is to be found, at least in part, in the following circumstances. A mathematical theory starts at ground level with the basic axioms, a set of propositions which are not to be questioned within the context of the theory. Then one deduces various conclusions from them; axioms 1 and 2 together imply conclusion A, axioms 2 and 3 imply conclusion B, conclusion A and axiom 4 imply conclusion C, and so on. From the ground level axioms one builds up a long chain of conclusions, and there seems to be no end to how far the chain can be extended. In other words, mathematical theories have a logical structure that is open at the top, closed at the bottom.

Scientists and others concerned with the real world live in a different logical environment. Nature does not reveal her "axioms" to us, and so we are obliged to start in the middle, at the level of our direct sense perceptions. From direct observations and their generalizations we can proceed upward, drawing arbitrarily long chains of conclusions, by reasoning which is perforce not as rigorous as that of Gauss and Cauchy, but with the compensation that our conclusions can be tested by observation. We do not proceed upward very far before the subject is called "engineering".

But from direct observations we also proceed downward, analyzing things into fundamentals, searching for deeper propositions from which all those above could have been deduced. This is called "pure science" and this chain of interlocking inferences can, as far as we know, also be extended indefinitely. So the game of science is, unlike that of mathematics, played on a logical field that is open at both the top and the bottom.

Consider now the two systems of probability theory created by mathematicians and scientists. The Kolmogorov system is a conventional mathematical theory, starting with ground axioms and building upward from them. If we start with probabilities of elementary events, such as "heads at each toss of a coin", we can proceed to deduce probabilities of more and more complicated events such as "not more than 137 or less than 93 occurrences of the sequence HHTTTHH in 981 tosses". In content, it resembles parts of the Bernoulli system, restated in set and measure–theory language.

But nothing in the Kolmogorov system tells us what probability should, in fact, be assigned to any real event. That is, it provides no rules for conversion of verbal information into probability assignments. Probabilities of the elementary events are simply given to us in the statement of a problem, as things determined elsewhere and not to be questioned. Probabilities of more complicated events follow from them by logical deduction following the postulated rules. Only one kind of probability (additive measure) exists, and the logical structure is closed at the bottom, open at the top. In this system, all is safe and certain and probabilities are absolute (indeed, so absolute that the very notion of conditional probability is awkward, unwanted, and avoided as long as possible).

In contrast, the Bayesian system of Laplace and Jeffreys starts with rules applied to some set of propositions of immediate interest to us. If by some means we assign probabilities to them, then we can build upward, introducing more propositions whose probabilities can be found as in the Kolmogorov system. The system is, as before, open at the top.

But it is not closed at the bottom, because in our system it is a platitude that all probabilities referring to the real world are, of necessity, conditional on our state of knowledge about the world; they cannot be merely postulated arbitrarily at the beginning of a problem. Converting prior information into prior probability assignments is an open–ended problem; you can always analyze further by going into deeper and deeper hypothesis spaces. So our system of probability is open at both the top and the bottom.

The downward analysis, that a scientist is obliged to carry out, represents for him fully half of probability theory, that is not present at all in the Kolmogorov system. Just for that reason, this neglected half is not as fully developed, and when we venture into it with a new application, we may find ourselves exploring new territory.

I think that most mathematicians are uncomfortable when the ground opens up and that safe, solid closed bottom is lost. But their security is bought only at the price of giving up contact with the real world, a price that scientists cannot pay. We are, of necessity, creatures of the bog.

Mathematicians sometimes dismiss our arguments as nonrigorous; but the reasoning of a physicist, engineer, biochemist, geologist, economist – or Sherlock Holmes – cannot be logical deduction because the necessary information is lacking. A mathematician's reasoning is no more rigorous than ours when he comes over and tries to play on our field. Indeed, it has been demonstrated many times that an experienced scientist could reason confidently to correct conclusions, where a mathematician was helpless because his tools did not fit the problem. Our reasoning has always been an intuitive version of Bayes' theorem, which can cope with incomplete information.

DOWNWARD ANALYSIS IN STATISTICAL MECHANICS

 15

To illustrate this open–ended descent into deeper and deeper hypothesis spaces, a physicist or chemist considering a common object, say a sugar cube, might analyze it into a succession of deeper and deeper hypothesis spaces on which probabilities might be defined. Our direct sense perceptions reveal only a white cube, with frosty rather than shiny sides, and no very definite hypothesis space suggests itself to us. But examination with a low power lens is sufficient to show that it is composed of small individual crystals. So our first hypothesis space H_1 might consist of enumerating all possible sizes and orientations of those crystals and assigning probabilities to them. Some of the properties of the sugar cube, such as its porosity, could be discussed successfully at that level.

On further analysis one finds that crystals are in turn composed of molecules regularly arranged. So a deeper hypothesis space H_2 is formed by enumerating all possible molecular arrangements. Before an x–ray structure determination is accomplished, our state of knowledge would be represented by a very broad probability distribution on H_2, with many arrangements "equally likely" and almost no useful predictions. After a successful structure analysis the "nominal" arrangement is known and it is assigned a much higher probability than any other; then in effect the revised probabilities on H_2 enumerate the possible departures from perfection, the lattice defects. At that level, one would be able to say something about other properties of sugar, such as cleavage and heat conductivity.

Then chemical analysis reveals that a sucrose molecule consists of 12 carbon, 22 hydrogen, and 11 oxygen atoms; a deeper hypothesis space H_3 might then enumerate their positions and velocities (the "phase space" of Maxwell and Gibbs). At this level, many properties of the sugar cube, such as its heat capacity, could be discussed with semiquantitative success; but full quantitative success would not be achieved with any probability distribution on that space.

Learning that atoms are in turn composed of electrons and nuclei suggests a still deeper space H_4 which enumerates all their possible positions and velocities. But as Arnold Sommerfeld found, H_4 leads us to worse inferences than H_3; fabulously wrong specific heats for metals. In this way, as in our first Sermon, Nature warns us that we are going in the wrong direction.

Still further analysis shows that the properties of atoms are only approximately, and those of electrons are not even approximately, describable in terms of positions and velocities. Rather, our next deeper hypothesis space H_5 is qualitatively different, consisting of the enumeration of their quantum states. This meets with such great success that we are still exploring it. In principle (*i.e.*, ignoring computational difficulties) it appears that all thermodynamic and chemical properties of sugar could be inferred quantitatively at that level.

Our present statistical mechanics stops at the level H_5 of enumerating the "global" quantum states of a macroscopic system. At that deepest level yet reached, simple counting of those states (multiplicity factors) is sufficient to predict all equilibrium macrostates; they are the ones with greatest multiplicity W (thus greatest entropy $\log W$) compatible with our macroscopic data. Thus while "equally likely" on H_2 had almost no predictive value, and "equally likely" on H_3 was only partially successful, "equally likely" on H_5 leads to what is probably the greatest predictive success yet achieved by any probabilistic theory.

Presumably, simple counting of quantum states will also suffice to predict all reproducible aspects of irreversible processes; but the computations are so huge that this area is still largely unexplored. We cannot, therefore, rule out the possibility that new surprises, and resulting further analysis, may reveal still deeper hypothesis spaces H_6 , H_7, and so on (hidden variables?). Indeed, the hope that this might happen has motivated much of the writer's work in this field.

But the fact that we do have great success with H_5 shows that still deeper spaces cannot have much influence on the predictions we are now making. As Henri Poincaré put it, rules which succeed " – will not cease to do so on the day when they become better understood." Even if we knew all about H_6, as long as our interest remained on the current predictions, we would have little to gain in pragmatic results, and probably much to lose in computational cost, by going to H_6. So, although in principle the downward analysis is open ended, in practice there is an art in knowing when to stop.

CURRENT PROBLEMS

In newer problems (image reconstruction, spectrum analysis, geophysical inverse problems, *etc.*, the analysis of deeper hypothesis spaces H_1, $H_2, \ldots$ is still underway, and we don't know how far it will go. It would be nice if we could go down to a space H_x deep enough so that on it some kind of "symmetry" points to what seem "equally possible" cases, with predictive value. Then probabilities on the space M of the observable macroscopic things would be just multiplicities on the deeper space H_x, and inference would reduce to maximizing entropy on H_x, subject to constraints specifying regions of M.

The program thus envisaged would be in very close analogy with thermodynamics. Some regard our efforts to cling to this analogy as quaint; in defense we note that statistical mechanics is the only example we have thus far of that deeper analysis actually carried through to a satisfactory stopping point, and it required over 100 years of effort to accomplish this. So we think that we had better learn as much as we can from this example.

But in the new problems we have not yet found any Liouville theorem to guide us to the appropriate hypothesis space, as Gibbs had to guide him to H_3 and was then generalized mathematically to H_5. For Gibbs, invariance of phase volume under the equations of motion and under canonical transformations – which he took great care to demonstrate and discuss at some length before entering into his thermodynamic applications – meant that assigning uniform prior probability, or weight, to equal phase volumes had the same meaning at all times and in all canonical coordinate systems. This was really applying the principle of group invariance, in just the way advocated much later by the writer (Jaynes, 1968).

Specifying our deepest hypothesis space, on which we assign uniform weight before adding any constraints to get the nonuniform MAXENT prior, is the means by which we define our starting point of complete ignorance but for enumeration of the possibilities, sometimes called "pre–prior" analysis. Long ago, Laplace noted this problem and stated that the exact appreciation of "equally possible" is "one of the most delicate points in probability theory". How right he was! Two hundred years later, we are still hung up on this "exact appreciation" in every new application.

For a time, writers thought they had evaded this delicate point by redefining a probability as a frequency; but in fact they had only restricted the range of applications of probability theory. For the general problems of inference now being attacked, the need to define what we mean by "complete ignorance" – complete, that is, but for enumeration of the possibilities to be considered in our problem – cannot be evaded, any more than the notion of zero could be evaded in arithmetic.

Today this is not just a puzzle for philosophers. It is crucially important that we learn how to build more prior information into our prior probabilities by developing that neglected half of probability theory. All inverse problems need this, and the possibility of any major progress in pattern recognition or artificial intelligence depends on it.

But in each area, pending a satisfactory analysis to a satisfactory stopping point, we can take some comfort in Tukey pragmatism (don't confuse the procedure with the reason):

> "A procedure does not have hypotheses. Rather, there are circumstances where it does better, and others where it does worse". ··· John W. Tukey (1980)

Our present maximum entropy procedure is supported by many different rationales, including:

(1) COMBINATORIAL Boltzmann, Darwin, Fowler

(2) INFORMATION THEORY — Shannon, Jaynes
(3) UTILITY — Good, Skilling
(4) LOGICAL CONSISTENCY — Shore, Johnson, Gull
(5) CODING THEORY — Rissanen
(6) PRAGMATIC SUCCESS — Gibbs, Papanicolaou, Mead

Of these, (1) is easy to explain to everybody, while (2) is more general, but hard to explain to those with orthodox statistical training, (3) and (4) are currently popular, and (5) shows great long–range theoretical promise, but is not yet well explored.

Most of the writer's recent discussions have concentrated on (1) rather than (2) in the belief that, after one has become comfortable with using an algorithm in cases where it has a justification so clear and simple that anyone can understand it, he will be more disposed to see a broader rationale for what he is doing.

It might be thought that, if many rationales all point to the same procedure, it is idle to argue about their relative merits. Indeed, if we were to stay forever on the current problems, different rationales would be just different personal tastes without real consequences. But different rationales generalize differently. In the current problems all these rationales happen to come together and point to the same procedure; but in other problems they would go their separate ways and point to different procedures. Therefore we think it is important in each application to understand the rationale and the circumstances as well as the procedure.

Of course, in case of doubt one can always fall back on (6). Doubtless, those who write the specific computer programs have done this a great deal, sometimes just trying out everything one can think of and seeing what works. We agree with Tukey that the theoretical justification of a procedure is often a mere tidying–up that takes place after the successful procedure has been found by intuitive trial and error.

But too much of that basically healthy Tukey pragmatism can lead one to take a negative view of theoretical efforts in general. The excessive disparagement of all theory, characteristic of that school, has been very costly to the field of data analysis; for Bayesian theory has a demonstrated ability to discover – in a few lines – powerful and useful procedures that decades of intuitive *adhockery* did not find.

We have already noted the writer's "Bayesian Spectrum and Chirp Analysis" given at the 1983 Laramie meeting on Maximum Entropy, where the Schuster periodogram acquires a new significance, leading to a very different way of using it in data analysis. Basically the same thing was noted by Steve Gull, who perceived the real Bayesian significance of the "dirty map" of radio astronomy (a two–dimensional analog of the periodogram), and therefore the proper way of using it in data analysis.

Other examples are Litterman's (1985) Bayesian economic forecasting method and the writer's Bayesian seasonal adjustment method noted above, both of which process the data in a way that takes into account previously neglected prior information.

G. E. P. Box (1982) also observes: "$\cdots$ history has shown that it is the *omission* in sampling theory, rather than the inclusion in Bayesian analysis, of an appropriate prior distribution, that leads to trouble."

In our next talk, "MONKEYS, KANGAROOS, AND N", we want to continue this line of thought, with more specific details about hypothesis spaces and rationales, for the particular case of image reconstruction. We want to make a start on the question whether some of that deeper analysis might have helped us. Our hypothesis spaces are still at the "Boltzmann level"; if we can understand exactly what is happening there, it might become evident that we need to go down at least to the "Gibbs level" and possibly beyond it, before finding a satisfactory stopping point for current problems.

REFERENCES

G. E. P. Box (1982) "An Apology for Ecumenism in Statistics", NRC Technical Report #2408, Mathematics Research Center, University of Wisconsin, Madison.

R. T. Cox (1946), "Probability, Frequency, and Reasonable Expectation", Am. Jour. Phys. **17**, 1–13. Expanded in *The Algebra of Probable Inference*, Johns Hopkins University Press, Baltimore (1961). Reviewed by E. T. Jaynes, Am. Jour. Phys. **31**, 66 (1963).

H. Cramér (1946), *Mathematical Methods of Statistics*, Princeton University Press.

J. Willard Gibbs (1875), "On the Equilibrium of Heterogeneous Substances" Reprinted in *The Scientific Papers of J. Willard Gibbs*, Vol. I, Longmans, Green & Co., 1906 and by Dover Publications, Inc., 1961.

J. Willard Gibbs (1902), *Elementary Principles in Statistical Mechanics*, Yale University Press, New Haven, Connecticut. Reprinted in *The Collected Works of J. Willard Gibbs*, Vol. 2, by Longmans, Green & Co., 1928 and by Dover Publications, Inc., New York, 1960.

E. T. Jaynes (1968), "Prior Probabilities", IEEE Trans. Systems Science and Cybernetics SSC–4, 227–241. Reprinted in V. M. Rao Tummala and R. C. Henshaw, eds, *Concepts and Applications of Modern Decision Models*, Michigan State University Business Studies Series, 1976; and in Jaynes (1983).

E. T. Jaynes (1976), "Confidence Intervals vs Bayesian Intervals", in W. L. Harper & C. A. Hooker, eds, *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, vol. II, Reidel Publishing Co., Dordrecht–Holland, pp. 175–257. Reprinted in Jaynes (1983).

E. T. Jaynes (1978) "Where do we Stand on Maximum Entropy?" in *The Maximum Entropy Formalism*, R. D. Levine & M. Tribus, eds, M.I.T. Press, Cambridge Mass., pp. 15–118. Reprinted in Jaynes, 1983).

E. T. Jaynes (1982), "On the Rationale of Maximum–Entropy Methods", Proc. IEEE **70**, pp 939–982.

E. T. Jaynes (1983) *Papers on Probability, Statistics, and Statistical Physics*, R. D. Rosenkrantz, ed., D. Reidel Publishing Co., Dordrecht–Holland.

H. Jeffreys (1939), *Theory of Probability*, Oxford University Press. Later editions, 1948, 1961, 1979.

P. S. Laplace (1812), *Théorie analytique des probabilités*, 2 vols. Reprints of this work are available from: Editions Culture et Civilisation, 115 Ave. Gabriel Lebron, 1160 Brussels, Belgium.

R. B. Litterman (1985), "Vector Autoregression for Macroeconomic Forecasting", in *Bayesian Inference and Decision Techniques*, A. Zellner & P. Goel, eds, North–Holland Publishers, Amsterdam.

E. C. Molina (1963), *Two Papers by Bayes, with Commentary*, Hafner Publishing Co., New York.

L. J. Savage (1954), *Foundations of Statistics*, J. Wiley & Sons. Second Revised Edition (1972) by Dover Publications, Inc., New York.

C. E. Shannon (1948), "A Mathematical Theory of Communication", Bell Systems Tech. Jour. **27**, 379, 623. Reprinted in C. E. Shannon & W. Weaver, *The Mathematical theory of Communication*, Univ. of Illinois Press, Urbana, 1949.

J. W. Tukey (1980), *The Practice of Spectrum Analysis*, University Associates, Princeton, N. J.

Notes for a special course given in December 1980.