THE RELATION OF BAYESIAN AND MAXIMUM ENTROPY METHODS

E. T. Jaynes
Arthur Holly Compton Laboratory of Physics
Washington University, St. Louis, Missouri 63130, U.S.A.

Abstract. Further progress in scientific inference must, in our view, come from some kind of unification of our present principles. As a prerequisite for this, we note briefly the great conceptual differences, and the equally great mathematical similarities, of Bayesian and Maximum Entropy methods.

We are all pleased at the progress that has been made, in many different fields, as a result of recent recognition of the power of Bayesian inference and the Maximum Entropy Principle (MAXENT). But this is not to say that further clarifications and technical developments aren't needed. It is a truism that every new level of understanding reached only reveals to us new questions of which we were unaware before.

Therefore, in spite of present successes, this is no time for relaxing our efforts to develop still better pragmatic algorithms and a more unified theoretical structure. Indeed, because of the pressure of new applications opened up by these very successes, the field of scientific inference has never been in greater need of new creative thought. But before we can hope to make much further progress, some clarification of our present principles is needed.

We have at present two principles, Bayes' theorem and MAXENT, that are held to have some fundamental status in the new domains. The practitioners of the art sometimes use one, sometimes the other; but beginners and critics alike seem puzzled by how we choose between them. How are these principles related to each other? Are they mutually consistent or in conflict? What is the proper place of each in our toolbox? Since nearly every conceivable opinion on these matters has been expressed already, one is hard put to say anything really new; but perhaps we may sift things out a bit.

At the most fundamental level as now perceived, by "applying Bayes' theorem" we mean *calculating* the probability

$$p(H|DI) = p(H|I)p(D|HI)/p(D|I) \qquad (1)$$

where, in our applications, H stands for some hypothesis whose truth we want to judge, D for a set of data, and I for whatever "prior information" we have in addition to the data. The prior probability p(H|I) of H gets updated to the posterior

G. J. Erickson and C. R. Smith (eds.),
Maximum-Entropy and Bayesian Methods in Science and Engineering (Vol. 1), 25–29.

probability p(H|DI) as a result of acquiring the data D. This includes parameter estimation, since H might be a statement about some property of a parameter θ.

By "applying MAXENT" we mean *assigning* a distribution $(p_1 \cdots p_n)$ on some "hypothesis space" by the criterion that it shall maximize the information entropy

$$S_I = -\sum p_i \log p_i \qquad (2)$$

subject to constraints that express properties we wish the distribution to have, but are not sufficient to determine it. Entropy is used as the criterion for resolving the ambiguity remaining when we have stated all the conditions we are aware of.

On the face of it, it is hard to imagine two procedures more different, mathematically or logically. Bayes' theorem expresses nothing more than that Aristotelian logic is commutative. The propositions

$$HD = \text{"H and D are both true"}$$

$$DH = \text{"D and H are both true"}$$

say the same thing, so they must have the same truth value and the same probability whatever our information I. Then in the product rule of probability we may interchange D and H:

$$p(DH|I) = p(D|HI)p(H|I) = p(H|DI)p(D|I) \qquad (3)$$

which is Bayes' theorem. Obviously, then, anyone who reasons in a way that conflicts with Bayes' theorem is violating a rather elementary principle of logic.

Fundamentally, a single application of Bayes' theorem gives us only a probability; not a probability distribution. Indeed, Bayes' theorem makes no reference to any sample space or hypothesis space; (H, D, I) may stand for any propositions with well-defined meanings. Just for that reason, Bayes' theorem cannot determine the numerical value of any probability directly from our information; to apply it one must first use some other principle to translate our information into numerical values for p(H|I), p(D|HI), p(D|I).

In scientific inference, therefore, before we can apply Bayes' theorem our problem must be developed beyond the "exploratory phase", to the point where it has enough structure to determine p(D|HI).

In contrast, MAXENT requires that we specify in advance a definite hypothesis space $H_1 \cdots H_n$ which sets down the possibilities to be considered. It gives us necessarily a probability distribution, not just a probability; it does not make sense to ask for the MAXENT probability of an isolated proposition H, that is not embedded in some hypothesis space of alternative propositions. But MAXENT does not require for input the numerical values of any probabilities on that space; rather it assigns those numerical values for us, directly out of our information, as expressed by our choice of hypothesis space and constraints. Therefore MAXENT can be applied in -- and is indeed most useful in -- the exploratory phase of a problem.

In these functional respects, MAXENT does for us almost the opposite of what Bayes' theorem does. How, then is it possible that two principles so different could be confused? This comes from two circumstances.

In the first place they have, after all, one feature in common; the updating of a state of knowledge. In MAXENT, for example, one may consider a problem with constraints X and Y, and find the solution $p_i(X,Y)$. Then a third constraint Z is added, and we re-maximize the entropy subject to all three constraints, leading to an updated solution $p_i(X,Y,Z)$. There is indeed a superficial resemblance to Bayes' theorem; and for some it requires only a sloppy notation and terminology -- calling these two MAXENT distributions "prior probabilities" and "posterior probabilities" -- to confuse them thoroughly.

Secondly, there is a technical circumstance which has caused trouble throughout the history of probability theory; different problems may lead to the same computational procedure. In some cases application of Bayes' theorem in one hypothesis space, and MAXENT in another, leads us to nearly identical calculations.

For example, starting with Darwin & Fowler in the 1920's, many have noted that the MAXENT procedure on the space S of a single trial, and the Bayes' theorem procedure on the extension space S^n of n trials are asymptotically equivalent as n becomes very large; the latter circumstance may be taken as the basis of the combinatorial rationale for MAXENT, which differs from the more fundamental probabilistic one noted above. Some other examples of this Bayes-MAXENT mathematical correspondence are given in Jaynes (1968). In a sense, this only illustrates their mutual consistency; but it can also be a rich source of confusion.

The recent literature has many attempts to clarify the relation of these principles. Williams (1980) sees Bayes' theorem as a special case of MAXENT, while van Campenhout & Cover (1981) see MAXENT as a special case of Bayes' theorem. In our view, both are correct as far as they go; but they consider only special cases. Zellner (1987) generalizes Williams' result.

Thus Williams considers the case where we have a set of possibilities $(H_1 \cdots H_n)$, and some new information E confines us to a subset of them. Such primitive information can be digested by either Bayes' theorem or MAXENT, leading of course to the same result; but Bayes' theorem is designed to cover far more general situations. Likewise, van Campenhout & Cover consider only the Darwin-Fowler scenario; MAXENT is designed to cover more general situations, where it does not make sense to speak of "trials".

Attempts to evade Bayes' theorem have been underway unceasingly since the rise of the "sampling theory" school of thought in the early 1900's. But we have already surveyed the results (Jaynes, 1983); whenever sampling theory methods have led us to different conclusions, closer examination has always shown the Bayesian results to be superior. Likewise, to the best of our knowledge, all attempts to extend Bayes' theorem [such as that of Jeffrey (1983)] have proved on closer examination to be satisfactory only in the cases where they agree with Bayes' theorem. Further strong evidence is given by Bretthorst (1987).

That MAXENT is in a similar position is indicated by the fact that the MAXENT procedure is in constant use, either as an analytical tool or as a computational algorithm, in a variety of very different problems; and no alternative has been found. Even those who reject the MAXENT principle are often led, by long and different reasoning, to the actual MAXENT algorithm and result.

To the best of our knowledge, all attempts to evade MAXENT in problems where we consider it appropriate, or to extend it to new problems, have been no more successful than the attempts to evade or extend Bayes' theorem. One does so only at the cost of getting results that can be shown to be defective or incomplete, in that they either fail to use all the relevant information or assume false information. We hope to discuss the evidence for this conclusion in much greater detail elsewhere.

There is indeed something fundamental about these principles, although we think that more unified ways of presenting them are still needed and will be found in the future. But we stress that our present principles and practice are fairly good; they have many demonstrable optimality properties and impressive pragmatic success. So unless we can recognize, and clearly understand the reason for, some specific defect in our present principles, we are hardly in a position to improve on them; as so much past experimence has shown, we are far more likely to lose some of the good performance features already accomplished.

An old adage among moralists is that "Virtue cannot be taught; only demonstrated"; and we must admit that today more worked-out examples of their analytical and numerical details, in a wider variety of real problems, are much needed to demonstrate how to apply them and what kind of results are to be expected. But with more and more books and Bayesian/MAXENT computer programs being written, this need should be filled soon.

REFERENCES

Bretthorst, L. (1987), Ph.D. Thesis, Department of Physics, Washington University, St. Louis, Missouri.

Jaynes, E. T. (1968), "Prior Probabilities", IEEE Trans. Systems Sci. Cybern. SSC-4, 227-241 (1968). Reprinted in V. M. Rao Tummala and R. C. Henshaw, Eds., Concepts and Applications of Modern Decision Methods (Michigan State University Business Studies Series, 1976), and in Jaynes (1983).

Jaynes, E. T. (1983), Papers on Probability, Statistics and Statistical Physics", R. D. Rosenkrantz, Editor, D. Reidel Pub. Co., Dordrecht-Holland. Reprints of 14 papers dated 1957-1980.

Jeffrey, R. C. (1983), The Logic of Decision, 2nd Edition, Univ. of Chicago Press.

van Campenhout, *J. & Cover, T. M. (1981), IEEE Trans. Inform. Theory IT-27,* 483.

Williams, P.M. (1980), "Bayesian Conditionalisation and the Principle of Minimum Information", Brit. J. Phil. Sci. 31, 131-144. The same mathematical fact was noted by R. D. Rosenkrantz, Inference, Method and Decision, D. Reidel, Dordrecht-Holland (1977); 55-57.

Zellner, A. (1987), "Optimal Information Processing and Bayes' Theorem" *The American Statistician* (to be published).