Optimal Information Processing and Bayes's Theorem

ARNOLD ZELLNER*

In this article statistical inference is viewed as information processing involving input information and output information. After introducing information measures for the input and output information, an information criterion functional is formulated and optimized to obtain an optimal information processing rule (IPR). For the particular information measures and criterion functional adopted, it is shown that Bayes's theorem is the optimal IPR. This optimal IPR is shown to be 100% efficient in the sense that its use leads to the output information being exactly equal to the given input information. Also, the analysis links Bayes's theorem to maximum-entropy considerations.

KEY WORDS: Information theory; Maximum entropy; Statistical inference.

1. INTRODUCTION

Bayes's theorem has been widely used as an inductive learning model to transform prior and sample information into posterior information in econometrics, statistics, physics, and other sciences. Although some works have been devoted to rationalizing Bayes's theorem as a coherent learning model (e.g., Cox 1961; Jaynes 1974, 1983, 1984; Jeffreys 1967, 1973), it does not appear that Bayes's rule has been derived as an optimal information processing rule. In what follows, this problem is addressed after describing some needed information measures and a criterion functional. This criterion functional is minimized in a calculus-of-variations approach to yield an operational, optimal information processing rule that is surprisingly identical to Bayes's rule. Further, Bayes's rule is informationally efficient in a sense defined in Section 2.

The plan of the article is as follows. In Section 2 needed concepts are introduced, the problem is explained, and its solution is presented. The derivation of the solution is given in Section 3. Some concluding remarks are provided in Section 4.

2. INFORMATION CONCEPTS AND AN OPTIMAL INFORMATION PROCESSING RULE

Let y denote the given data, and let $l(\theta|y)$ denote a likelihood function for θ—a parameter, scalar, or vector contained in the parameter space Θ. Further, let

$\pi_a(\theta|I) \equiv$ given prior or antedata probability density function (pdf) for $\theta \subset \Theta$, based on prior information I,

$\pi_a(\theta|I) \equiv$ given prior or antedata probability density function (pdf) for $\theta \subset \Theta$, based on prior information I,

$\pi_p(\theta|D) \equiv$ postdata pdf for $\theta \subset \Theta$, where $D = (y, I)$, the given sample, y, and prior information, I,

$p(y|I) \equiv$ pdf for y, given by

$$p(y|I) \equiv \int_\theta \pi_a(\theta|I) f(y|\theta)d\theta, \tag{2.1}$$

where $f(y|\theta)$ is the pdf for y given θ. [The term *postdata pdf* is employed instead of *posterior pdf* to emphasize that the optimal form of $\pi_p(\theta|D)$ is to be derived.]

Note that (2.1) is a definition that does not involve the assumption that the product rule of probability theory necessarily holds. See Jeffreys (1967, pp. 25, 52) for a discussion of assumptions needed for the product rule to be valid.

The inputs and outputs of any information processing rule (IPR) are depicted graphically in Figure 1, where $l(\theta|y)$, the likelihood function, is $f(y|\theta)$ viewed as a function of θ. Thus the information in the likelihood function $l(\theta|y)$ and the prior pdf $\pi_a(\theta|I)$ enter the IPR, whose output is the information in the postdata pdf $\pi_p(\theta|D)$ and the pdf $p(y|I)$. Different IPR's will produce different output information from the given input information. Some IPR's may be *inefficient* in the sense that the output information, measured in a suitable metric, is *less* than the input information. On the other hand, some IPR's may add *extraneous* information so that the output information is greater than the given input information, an undesirable state of affairs. A good, efficient IPR will satisfy the following principle.

Information Conservation Principle (ICP). Input information = Output information.

An IPR that satisfies the ICP is 100% efficient in the sense that the ratio of output to input information is equal to 1. An inefficient IPR has efficiency less than 100%. An IPR that *adds* extraneous information is considered unsatisfactory.

To implement these concepts, there is a need to measure the information in the input and output pdf's. (For discussion of information measures, see Kullback 1959.) The following postdata measures will be employed:

$$\text{Information in } l(\theta|y) \equiv \int_\theta \pi_p(\theta|D)\log l(\theta|y)d\theta. \tag{2.2}$$

$$\text{Information in } \pi_a(\theta|I) \equiv \int_\theta \pi_p(\theta|D)\log \pi_a(\theta|I)d\theta. \tag{2.3}$$

$$\text{Information in } \pi_p(\theta|D) \equiv \int_\theta \pi_p(\theta|D)\log \pi_p(\theta|D)d\theta. \tag{2.4}$$

$$\text{Information in } p(y|I) \equiv \int_\theta \pi_p(\theta|D)\log p(y|I)d\theta = \log p(y|I). \tag{2.5}$$

In each case, information is given as an average of a log pdf with $\pi_p(\theta|D)$ used as a "weight function."

To illustrate some of these measures, suppose $l(\theta|y) = (2\pi/n)^{-1/2} \times \exp\{-n(\bar{y} - \theta)^2/2\}$, where $\bar{y}$ = sample mean. Then $\log l(\theta|y) = -1/2 \log 2\pi/n - n(\bar{y} - \theta)^2/2$, and (2.2) yields $-1/2 \log 2\pi/n - n[\mathrm{var}(\theta|D) + (\bar{y} - \bar{\theta})^2]/2$, where $\bar{\theta}$ and $\mathrm{var}(\theta|D)$ are the mean and variance of

*Arnold Zellner is Professor of Economics and Statistics, Graduate School of Business, University of Chicago, Chicago, Illinois 60637. The research for this article was financed in part by the National Science Foundation and the H. G. B. Alexander Endowment Fund of the University of Chicago's Graduate School of Business.

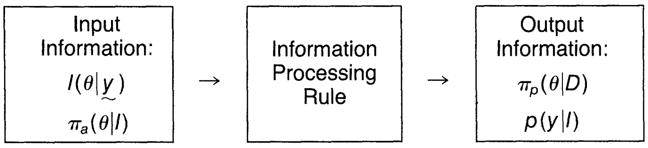

Figure 1. Inputs and Outputs of Any Information Processing Rule.

$\pi_p(\theta|D)$, respectively. The larger $\mathrm{var}(\theta|D)$ and $(\bar{y} - \bar{\theta})^2$ are, the smaller the information in $l(\theta|y)$ is, which is reasonable. Further, if $\pi_a(\theta|I) = (2\pi)^{-1/2} \exp\{-(\theta - m)^2/2\}$, where m is a given prior mean, $\log \pi_a(\theta|I) = -1/2 \log 2\pi - (\theta - m)^2/2$ and (2.3) yields $-1/2 \log 2\pi - [\mathrm{var}(\theta|D) + (m - \bar{\theta})^2]/2$; thus the larger $\mathrm{var}(\theta|D)$ and $(m - \bar{\theta})^2$ are, the smaller the information in the pdf $\pi_a(\theta|I)$ is. As regards (2.4), if $\pi_p(\theta|D) = (2\pi)^{-1/2} \times \exp\{-(\theta - \bar{\theta})^2/2\}$, the information in this pdf is $-1/2 \log 2\pi - \mathrm{var}(\theta|D)/2$, again a reasonable result because the information diminishes as $\mathrm{var}(\theta|D)$ grows larger.

Using the information measures (2.2)–(2.5), the problem is to determine the form of $\pi_p(\theta|D)$, given the inputs $l(\theta|y)$ and $\pi_a(\theta|I)$, so as to minimize a reasonable criterion functional subject to the condition that $\pi_p(\theta|D)$ is a proper pdf. The criterion functional that will be employed is motivated by the ICP. That is, $\pi_p(\theta|D)$'s form should be such that the output information is as close as possible to the input information and, ideally, equal to it. Thus the criterion functional

$$\begin{aligned}\Delta[\pi_p(\theta|D)] &= \int_\theta \pi_p(\theta|D) \log \pi_p(\theta|D)d\theta + \log p(y|I) \\ &\quad - \int_\theta \pi_p(\theta|D)[\log l(\theta|y) \\ &\quad + \log \pi_a(\theta|I)]d\theta \qquad (2.6)\end{aligned}$$

will be minimized to $\int_\theta \pi_p(\theta|D)d\theta = 1$. Note from (2.2)–(2.5) that the first two terms on the right side of (2.6) represent the information in the outputs, $\pi_p(\theta|D)$ and $p(y|I)$, and from this is subtracted the information in the inputs, $l(\theta|y)$ and $\pi_a(\theta|I)$—the likelihood function and the prior pdf, respectively. Thus (2.6) represents the difference between the output and input information, and minimization with respect to the choice of $\pi_p(\theta|D)$, the postdata pdf, will make the output information as close as possible to the input information.

It is useful to note that the criterion functional in (2.6) can be expressed as

$$\Delta[\pi_p(\theta|D)] = 2\int_\theta \pi_p(\theta|D) \log \left[\frac{\pi_p(\theta|D)}{\pi_a(\theta|I)} \times \frac{p(y|I)}{f(y|\theta)}\right]^{1/2} d\theta. \qquad (2.7)$$

From (2.7), it is seen that minimizing (2.6) involves choosing $\pi_p(\theta|D)$ such that the postdata mean of the logarithm of the geometric mean of the ratios $\pi_p(\theta|D)/\pi_a(\theta|I)$ and $p(y|I)/f(y|\theta)$ will be as small as possible; in this sense the outputs, $\pi_p(\theta|D)$ and $p(y|I)$, will be as "close" as possible to the inputs, $\pi_a(\theta|I)$ and $f(y|\theta)$. Equation (2.7) can be interpreted as an information-theory divergence measure relating to the pdf's $\pi_a(\theta|y, I)p(y|I)$ and $\pi_a(\theta|I)f(y|\theta)$ and, as a referee suggested, the negative entropy of $\pi_p(\theta|D)$ relative to the measure $\pi_a(\theta|I)f(y|\theta)/p(y|I)$.

As shown in Section 3, the solution, denoted by $\pi^*(\theta|D)$, to the minimization problem is

$$\pi_p^*(\theta|D) = c\pi_a(\theta|I)l(\theta|y), \qquad (2.8)$$

with $c^{-1} = \int \pi_a(\theta|I)l(\theta|y)d\theta = p(y|I)$. From (2.8), it is seen that $\pi_p^*(\theta|D)$ is just the postdata or posterior pdf yielded by Bayes's IPR, that is, Bayes's theorem. From (2.8), $\int_\theta \pi_p^*(\theta|D) \log[\pi_p^*(\theta|D)/\pi_a(\theta|I)]d\theta = \int_\theta \pi_p^*(\theta|D) \log l(\theta|y) d\theta - \log p(y|I)$, where the quantity on the left side is the negative of the entropy of $\pi_p^*(\theta|D)$ relative to the measure $\pi_a(\theta|I)$. Thus the negative entropy of the posterior pdf can be expressed in terms of the information measures.

To check the informational efficiency of the rule in (2.8), substitute $\pi_p^*(\theta|D)$ given in (2.8), with $c = 1/p(y|I)$, into (2.6) with the result

$$\Delta[\pi_p^*(\theta|D)] = 0. \qquad (2.9)$$

From (2.9), the IPR in (2.8) is 100% efficient and therefore satisfies the ICP relative to the information measures in (2.2)–(2.5). That is, use of $\pi_p^*(\theta|D)$ as a postdata pdf makes the input information equal to the output information. No information is lost and no extraneous information is introduced by use of the Bayesian IPR in (2.8).

3. DERIVATION OF THE OPTIMAL INFORMATION PROCESSING RULE

To minimize the criterion functional $\Delta[\pi_p(\theta|D)]$ in (2.6) subject to the condition that $\pi_p(\theta|D)$ be a proper, normalized pdf, we consider the class of neighboring functions, $\bar{\pi}_p(\theta|D) = \pi_p(\theta|D) + \varepsilon\eta(\theta)$, where ε is a small quantity and $\eta(\theta)$ is an arbitrary continuous function with a value of 0 at the endpoints of the region of integration and with $\int_\theta \eta(\theta)^2 d\theta < \infty$. On substituting $\bar{\pi}_p(\theta|D)$ in (2.6) and in the side condition $\int \pi_p(\theta|D)d\theta = 1$, the Lagrangian expression, denoted by $L(\varepsilon)$, is

$$\begin{aligned}L(\varepsilon) &= \int_\theta [\pi_p(\theta|D) + \varepsilon\nu(\theta)] \log[\pi_p(\theta|D) + \varepsilon\nu(\theta)]d\theta \\ &\quad - \int [\pi_p(\theta|D) + \varepsilon\nu(\theta)] \log[\pi_a(\theta|I) + \log l(\theta|y)] \\ &\quad + \log p(y|I) + \lambda[\int[\pi_p(\theta|D) + \varepsilon\nu(\theta)]d\theta - 1], \\ &\qquad (3.1)\end{aligned}$$

where λ is a Lagrange multiplier. On differentiating $L(\varepsilon)$ with respect to ε and evaluating the derivative at $\varepsilon = 0$, the necessary condition for an extremum is

$$\begin{aligned}L'(0) &= \int_\theta \nu(\theta) \, [\log \pi_p(\theta|D) + 1 - \log \pi_a(\theta|I) \\ &\quad - \log l(\theta|\underset{\sim}{y}) + \lambda]d\theta = 0.\end{aligned}$$

For $L'(0)$ to be equal to 0 for any arbitrary $\nu(\theta)$, the quantity in brackets in the integrand must be identically equal to 0, which leads to

$$\pi_p(\theta|D) = \pi_p^*(\theta|D) = c\pi_a(\theta|I)l(\theta|y), \qquad (3.2)$$

where $c = e^{-(1+\lambda)}$ is given by $c^{-1} = \int_\theta \pi_a(\theta|I)l(\theta|y)d\theta = p(y|I)$. Further, $d^2L(\varepsilon)/d\varepsilon^2$, evaluated at $\varepsilon = 0$, is given by

$$d^2L(\varepsilon)/d\varepsilon^2\big|_{\varepsilon=0} = \int_\theta \nu(\theta)^2/\pi_p^*(\theta|D)d\theta. \qquad (3.3)$$

Assuming $\pi_p^*(\theta|D) \leq M$, a positive constant, (3.3) is larger than $(1/M) \times \int_\theta \nu(\theta)^2 d\theta > 0$ and the expression for $\pi_p(\theta|D)$ in (3.2) corresponds to a minimum. It is assumed that the integral in (3.3) converges as would be the case of $\lim_{\theta \to \infty} |\theta|^q [\nu(\theta)^2 / \pi_p^*(\theta)] \to$ constant for $q > 1$. This requirement places a condition on the rate at which $\nu(\theta)^2 \to 0$ as $\theta \to \pm\infty$. Alternatively, if the parameter space is finite, the integral will usually converge.

Thus it is seen that (3.2), in the form of Bayes's theorem, is a solution to the constrained minimization problem. Further, when $\pi_p^*(\theta|D)$ in (3.2) is inserted in (2.6), the result is $\Delta[\pi_p^*(\theta|D)] = 0$; therefore, (3.2) is informationally 100% efficient.

4. CONCLUDING REMARKS

In this article an information processing approach has been formulated. This approach is thought to be a useful representation of the processing of information in inference situations. An optimal information processing rule was derived that is identical to Bayes's rule and is 100% informationally efficient. Further research to consider extended variants of the criterion functional used in this study as well as alternative measures of information would be valuable.

[*Received October 1986. Revised July 1987.*]

REFERENCES

Cox, R. T. (1961), *The Algebra of Probable Inference*, Baltimore: Johns Hopkins University Press.

Jaynes, E. T. (1974), "Probability Theory With Application in Science and Engineering: A Series of Informal Lectures," unpublished manuscript, Washington University, Dept. of Physics.

——— (1983), *Papers on Probability, Statistics and Statistical Physics*, ed. R. D. Rosenkrantz, Dordrecht, Holland: D. Reidel.

——— (1984), "The Intuitive Inadequacy of Classical Statistics," *Epistemologica*, 7, 43–74.

Jeffreys, H. (1967), *Theory of Probability* (3rd ed.), London: Oxford University Press.

——— (1973), *Scientific Inference* (3rd ed.), Cambridge, U.K.: Cambridge University Press.

Kullback, S. (1959), *Information Theory and Statistics*, New York: John Wiley.

Comment

E. T. JAYNES*

Arnold Zellner's article appears to me a potentially important one in two respects, one psychological and one theoretical. By looking at Bayes's theorem in a fresh way independent of previous arguments, it could make the use of Bayesian methods more attractive and widespread, and stimulate new developments in the general theory of inference.

In almost all real problems of scientific inference we need to take into account our total state of knowledge, only part (or sometimes none) of which consists of frequencies. For many years I have believed, and taught my students, that the fundamental justification for use of Bayes's theorem in such applications lies in the logical consistency arguments of R. T. Cox, referred to by Zellner. Cox's desiderata appeared to me more elementary—therefore, more compelling logically—than the well-known arguments of De Finetti, Jeffreys, and L. J. Savage.

Recently, however, I was taken aback in a conversation with a prominent anti-Bayesian when he opined that logical consistency is not an important desideratum at all for inference, because it gives no reason to believe that our conclusions are in any way sensible from a pragmatic standpoint.

It appears, then, that the arguments that convinced me may have little psychological force for others; perhaps this may account for the rather slow growth of Bayesian methods, in spite of their easy success in applications where sampling-theory methods would be awkward due to nuisance parameters, nonexistence of sufficient or ancillary statistics, or cogent prior information calling for an informative prior. It is surely clear to all, however, that inference is basically a procedure of information processing; some black box receives input information in the form of prior knowledge and data, and it emits output information in the form of parameter estimates, predictive distributions, and so forth.

Then a derivation of Bayes's theorem directly from desiderata of optimal information processing might have a stronger convincing power for many. An acceptable inference procedure should have the property that it neither ignores any of the input information nor injects any false information; if this requirement already determines Bayes's theorem, the issue would seem to be settled.

The logarithmic measures of information might appear arbitrary at first glance; yet as Kullback showed, this is not the case. And Bayes's theorem doubtless has more than one information-optimality property; I rather expect that, having seen this start, others different in detail and/or background conditions may be found. Indeed, the fact that many different psychological approaches point to the same actual algorithm is a major strength of "Bayesianity."

On the theoretical side, entropy has been a recognized part of probability theory since the work of Shannon 40 years ago, and the usefulness of entropy maximization as a tool in generating probability distribution is thoroughly established in numerous new applications including statistical mechanics, spectrum analysis, image reconstruction, and biological macromolecular structure determination. Grandy

*E. T. Jaynes is Wayman Crow Professor of Physics, Department of Physics, Washington University, St. Louis, Missouri 63130.

