A RECYCLING OF THE STATISTICAL PRINCIPLE OF MAXIMUM INFORMATION (*)

by ANDREA SGARRO (in Trieste) (**)

SOMMARIO. - *Si è detto che il principio di massima informazione è un criterio oggettivo per scegliere la distribuzione di probabilità iniziale nell'inferenza statistica di tipo neo-bayesiano. In questo articolo si vuol mostrare come l'uso di tale principio non sia coerente, a meno che lo sperimentatore non si renda conto che si sottace un'ipotesi, la cui accettazione dipende solo dalla sua scelta soggettiva.*

SUMMARY. - *The principle of maximum information has been described as an objective criterion for choosing the initial probability distribution in Bayesian statistical inference. In this paper we argue that the use of the principle is inconsistent, unless the experimenter realizes that an assumption is tacitly understood, which depends only on his subjective choice.*

1. Introduction.

The principle of maximum information has been described as an objective criterion for choosing the initial probability distribution in Bayesian statistical inference (the initial probability distribution expresses the beliefs of the experimeter about a parameter before the experiments are actually carried out). In loose terms the principle reads: if the initial probability distribution is constrained to belong to a set F and is otherwise arbitrary, choose the probability distribution of F whose entropy is a maximum.

(*) Pervenuto in Redazione il 29 novembre 1978.
Lavoro eseguito nell'ambito del Gruppo Nazionale per l'Informatica Matematica del C. N. R..

(**) Indirizzo dell'Autore: Istituto di Matematica dell'Università - Piazzale Europa 1 - 34100 Trieste.

The heuristic justification is that: i) entropy is an adequate measure of uncertainty, ii) when the experimenter does not know anything about a parameter, except that its probability distribution must belong to *F*, it is natural that he chooses the probability distribution of *F* which corresponds to maximum uncertainty. The use of the principle is supported by statisticians who, though not disregarding the value of prior information, believe that the choice of the initial probability distribution must be « objective » (they oppose any subjective element in statistical inference).

As a matter of fact the principle of maximum information is an extension of the famous « principle of insufficient reason », due to Laplace, which postulates the choice of the uniform distribution when there are no constraints, that is when *F* is the set of all probability distributions over the parameter space.

Serious objections can be raised against this pretence of objectivization. Suffice it here to recall that the development of Bayesian inference has been much hampered precisely by the weakness of Laplace's principle.

Moreover, entropy is an adequate measure of uncertainty only for discrete probability distributions, while the principle has been applied even in the continuous case.

In this paper we use also the principle of « minimum difference of information » which is equally motivated both in the discrete and in the continuous case. A probability distribution *P* is given and the experimenter has to choose, out of a set *F*, that probability distribution which « resembles » *P* as much as possible; to put it differently, the experimenter has to « adapt » *P* to belong to the constrained set *F*. As a measure of « dissimilarity » between probability distributions, the informational divergence (Kullback's discrimination) is chosen. The principle of maximum information, both in its discrete and continuous version, is formally re-obtained when *P* is the uniform distribution.

By giving a suitable conceptual interpretation of the principle of minimum difference of information, we shall show in the body of the paper that the use of the principle of maximum information, at least in its original « objective » version, is inconsistent: in fact an assumption is tacitly understood, which depends on a subjective choice of the experimenter.

The following three sections are devoted to some technical preliminaries on entropy and divergence (which the experienced reader can skip), to the principle of maximum information and its shortcomings, and to a discussion in which we make use of the principle of minimum difference of information to support our arguments. Technicalities have been avoided whenever possible.

2. Entropy and Informational Divergence.

It is a difficult task to summarize the results about entropy and divergence [1] which are scattered in the literature; the reader is referred, e. g., to Guiaşu's book [1] which contains a reasonably extensive account of the many-sided problems involved (information-theoretic, statistical, etc.). Here we shall only recall some very basic points which we need in the sequel. First, let us consider the finite case.

Assume A is a finite set of k elements. Then the probability distributions (p. d.'s) over A can be identified with the probability vectors of length k: $P=(p_1, \dots, p_k)$, $p_j \geq 0$, $1 \leq j \leq k$, $\Sigma p_j = 1$ (unspecified summations are meant over the whole set of indices); p_j is the probability of the *j-th* element of A.

The entropy $H(P)$ is defined as:

$$H(P) = -\Sigma p_j \log p_j; \tag{1}$$

if $Q=(q_1, \dots, q_k)$ is also a p. d., the divergence $D(P; Q)$ of P from Q is defined as:

$$D(P; Q) = \Sigma p_j \log (p_j/q_j) \tag{2}$$

($0 \log 0$ is to be interpreted as 0; $D(P; Q)$ is infinite when P is not absolutely continuous with respect to Q).

When Q is the uniform p. d., (2) becomes:

$$D(P; Q) + H(P) = \log k = \text{const.} \tag{3}$$

Therefore, in this case, minimizing $D(P; Q)$ with respect to P is the same as maximizing $H(P)$.

$H(P)$ and $D(P; Q)$ are non-negative quantities; $H(P)$ ranges from 0 (when P has a 1 among its components) to $\log k$ (when P is the uniform probability distribution); $D(P; Q)=0$ iff $P=Q$.

[1] The (informational) divergence has been called also discrimination, or Kullback's number, or relative entropy.

There is large evidence in the literature which suggests the following heuristic interpretations:

i) $H(P)$ is an adequate measure of the « uncertainty » contained in the p. d. P;

ii) $D(P; Q)$ is an adequate measure of how much the p. d. P « differs » from the p. d. Q [2].

Instead of A take now the real line R. For simplicity we shall confine ourselves to p. d.'s defined through density functions $f(x): f(x) \geq 0$, $x \in R$, $\int f(x)\,dx = 1$ (unspecified integrals are meant over R; readers familiar with measure theory will have no difficulty in extending properly our arguments).

If the p. d.'s P and Q are defined by the density functions $f(x)$ and $g(x)$, respectively, definitions (1) and (2) become:

$$H(P) = -\int f(x) \log f(x)\,dx, \tag{4}$$

$$D(P; Q) = \int f(x) \log \frac{f(x)}{g(x)}\,dx.$$

(Set $D(P; Q) = +\infty$ when P is not absolutely continuous with respect to Q).

If Q is the uniform distribution over the interval $[a, b]$, and if the density function of P is constrained to be null outside $[a, b]$, the following relation holds:

$$D(P; Q) + H(P) = \log(b-a) = \text{const.} \tag{5}$$

Therefore, in this case, minimizing $D(P; Q)$ with respect to P is the same as maximizing the entropy $H(P)$.

Unfortunately, while the properties of divergence which lead to the heuristic interpretation ii) are valid also in the continuous case, this is no longer true for the continuous entropy, as defined in (4) [3]. However the observation which follows (5) will prove useful.

[2] Note, however, that $D(P; Q)$ is *not* a metric in the topological sense of the word; $D(P; Q)$ is not symmetric and it must be viewed as a «distance from» rather than a « distance between »; cf. [2] and [3].

[3] The continuous entropy is not even necessarily positive.

3. The Principle of Maximum Information.

The principle of maximum information has been introduced, independently, by S. Kullback and R. A. Leibler (1951), E. T. Jaynes (1957) and R. S. Ingarden (1963). Its use is supported by statisticians who believe in an « objective » Bayesian approach to statistical inference. Their views will be made clear by the following quotation from [4] (passim): « ... it appears that statistical practise has reached a level where the problem of prior probabilities can no longer be ignored or belittled. The « personalistic » school of thought recognizes this fact, but proceeds to overcompensate it by offering us many different priors for a given state of prior knowledge. Surely, the most elementary requirement of consistency demands that two persons with the same relevant prior information should assign the same prior probabilities. Personalistic doctrine makes no attempt to meet this requirement. An unfortunate impression has been created that rejection of personalistic probabilities automatically means the rejection of Bayesian methods in general. This is not the case; the problem of achieving objectivity for prior probability assignments is not one of psychology or philosophy, but one of proper definitions and mathematical techniques. Prior probabilities can be made fully « objective » ».

Let us now illustrate the principle of maximum information.

Let F (the *constrained set*) be a set of probability vectors over A. Assume an experimenter has to choose a probability vector over A to express his prior information about a parameter which will be subsequently the object of a statistical investigation. Before starting the investigation the experimenter's knowledge of the parameter is limited to the fact that the probability vector to choose is constrained to belong to F; he lacks any other information. The principle of maximum information reads: in this case choose that probability vector in F which maximizes $H(P)$ [4].

The principle is currently justified by referring to the heuristic interpretation i) of entropy. The very serious objections which can be raised against the principle is that the notion of lack of information is fuzzy and, as it has been repeatedly shown, leads easily to contradictions.

[4] Here and in the sequel we shall assume that all maximization and minimization problems have a unique solution; for regularity conditions see [2].

EXAMPLE 1. Take a coin and assume three possibilities: « head », « tail » and « the coin stands upright ». It is reasonable to say that the entropy-maximizing distribution $\{1/3, 1/3, 1/3\}$ reflects maximum uncertainty as to the possible outcome; but it is questionable whether it reflects maximum lack of information on the side of the experimenter who is going to flip the coin. There seems to be some semantic confusion between the notions of « uncertainty » and « ignorance ». Only the first allows of a formalization, at least in the discrete case.

EXAMPLE 2. Assume F is the set of all real p. d.'s which are null outside $[0, 1]$ with probability 1. Then the entropy-maximizing p. d. for the parameter ϑ is uniform. Since the experimenter lacks any information also about the parameter $\mu = \sqrt{\vartheta}$, μ should be itself uniformly distributed over $[0, 1]$, according to the principle of maximum information. But this is a contradiction, since uniform distribution for ϑ implies a non-uniform distribution for $\sqrt{\vartheta}$.

Moreover, since the heuristic interpretation i) of entropy is not generally valid, it would seem that there is no hope of extending the principle of maximum information to the continuous case. Consider however the following well-known result (a proof can be found, e. g., in [1]): if F contains all the p. d.'s with fixed mean and variance, $H(P)$ is maximized by the normal distribution. If one has in mind the role of the normal distribution, it is difficult to resist the temptation of interpreting the theorem by saying that the normal distribution contains the largest amount of uncertainty compatible to given mean and variance; this, however, contradicts our statement that the entropy is not an adequate measure of uncertainty in the continuous case.

The shortcomings of the principle of maximum information will be discussed in the next section.

4. The Principle of Minimum Difference of Information.

The principle of minimum difference of information has been summarily described in section 1. Here we are going to introduce it in a somewhat different conceptual way.

Let us go back to the situation when an experimenter has to choose a p. d. to describe his prior information about a parameter. We shall give up any pretence of formalizing « complete » or « partial ignorance », or anything of the sort. Before starting his statistical investigation the experimenter has his own, possibly vague, beliefs: let him express them by choosing accordingly a p. d. Q. This is exactly what the « personalistic » school recommands: no easy recipe is given to make the choice.

Assume, however, that *later* (but before starting the statistical investigation) the experimenter learns that the prior p. d. must belong to a certain set F of p. d.'s. Instead of choosing a prior p. d. afresh, he might prefer to « adapt » his first choice Q, for example by choosing the p. d. P in F which minimizes $D(P; Q)$. This is reasonable, since the divergence is an adequate measure of dissimilarity between p. d.'s, both in the continuous and in the discrete case.

Note that the above procedure is somehow arbitrary: in fact the divergence is not the only adequate measure of dissimilarity between p. d.'s; moreover it is not clear why Q should appear as the second argument of $D(\cdot;\cdot)$ (recall that the divergence is not symmetric). Thus the choice of $D(\cdot; Q)$ as an adequate measure of dissimilarity from Q is itself subjective, or « personalistic ». We stress however that the principle of minimum difference of information *is not inconsistent,* once its subjective character is realized.

On the other hand our aim is *not* to support the use of the principle of minimum difference of information in statistical practice; rather we want to investigate the real meaning of the principle of maximum information. Now we see that the latter is re-obtained as a *particular case of the principle of minimum difference of information when the first choice p. d. Q is uniform over the parameter space*; in fact in this case minimizing the divergence $D(P; Q)$ in F is the same as maximizing the entropy $H(P)$ in the same set (cf. (3) and (5)).

A clarification is needed when the parameter space is such that the uniform p. d. over it is improper [5], for example in the case of the whole real line R. Let us go back to (5), with $[-a, a]$ instead of $[a, b]$. When $a = +\infty$, then uniform p. d. is improper and $D(P; Q)$ is undefined. However (5) provides a sound heuristical justification for the following rule: when Q is the improper uniform distribution over R, instead of minimizing $D(P; Q)$, maximize $H(P) = -\int f(x) \log f(x)\, dx$. Therefore our interpretation of the result at the end of section 3 is the following: of all the p. d.'s with given mean and variance, the normal

[5] We do not intend to give a rigorous definition of improper distributions; we recall however that the improper p. d. over R induces « proper » uniform p. d.'s over the intervals of R.

distribution is the « nearest » to the improper uniform distribution over R, in the sense of the principle of minimum difference of information.

A consistent experimenter uses the principle of maximum information only when he would have chosen (subjectively) the uniform p. d. over the parameter space, had the constrained set F contained all the p. d.'s over that space. There are very common cases when this does not happen: for example, in Bayesian inference for normal populations « complete ignorance » about the variance σ^2 is often expressed by assuming an improper uniform distribution for $\log \sigma^2$, so that the distribution of σ^2 is *not* uniform.

As a conclusive remark, we observe that, in order to get rid of the inconsistencies of the principle of maximum information in its « objective » version, we had to adopt a « personalistic » approach to statistical inference.

REFERENCES

[1] S. Guiașu: *Information Theory with Applications.* Mc Graw - Hill, 1977.

[2] I. Csiszár: *I-divergence Geometry of Probability Distributions and Minimization Problems.* The Annals of Probability (1) 3 (1975), 146-158.

[3] A. Sgarro: *An Informational Divergence Geometry for Stochastic Matrices.* Calcolo, (15) 1 (1978), 41-49.

[4] E. T. Jaynes: *Prior Probabilities.* IEEE Trans. on Systems Science and Cybernetics (SSC - 4) 3 (1968), Special Issue on Decision Analysis, 227-240.