Brit. J. Phil. Sci. 31 (1980), 131–144 *Printed in Great Britain*

Bayesian Conditionalisation and the Principle of Minimum Information

by P. M. WILLIAMS

The use of the principle of minimum information, or equivalently the principle of maximum entropy, has been advocated by a number of authors over recent years both in statistical physics as well as more generally in statistical inference.[1] It has perhaps not been sufficiently appreciated by philosophers, however, that this principle, when properly understood, affords a rule of inductive inference of the widest generality.[2] The purpose of this paper is to draw attention to the generality of the principle. Thus the Bayesian rule of conditionalisation, as well as its extension by R. C. Jeffrey, will be exhibited as special cases. General conditions under which it yields a unique prescription will also be studied. Detailed treatment will be restricted to the finite-dimensional case but an outline of the general case is given in the Appendix.

The underlying idea of maximum entropy inference is this. Suppose P to be a probability distribution assigning probabilities $p_1, \ldots, p_n$ to n mutually exclusive and jointly exhaustive events. Then the information-theoretic *entropy* of the distribution is defined by

$$S(P) = -\sum_{j=1}^{n} p_j \log p_j,$$

where it is to be understood that any term in the summation for which $p_j = 0$ vanishes. $S(P)$ can be thought of as measuring the 'missing information' or 'uncertainty' associated with the distribution. In particular, $S(P)$ assumes its maximum value for the uniform distribution and its minimum value for any distribution concentrated on a single event. The *principle of maximum entropy* states that the probability distribution appropriate to a given state of information is one that maximises entropy subject to the constraints imposed by the information given.

Suppose it to be a question of assigning probabilities to the outcomes of the throw of a loaded die where, specifically, the information available

[1] See, for example, Jaynes [1957] and Kullback [1959]. An extensive bibliography may be found in Guiaşu [1977].

[2] By a rule of inductive inference is meant a rule for passing from a prior probability distribution to a posterior distribution in the light of new information. The choice of an initial prior is a separate question. On this see, for example, Jaynes [1968].

Received 15 *December* 1978

makes it appropriate to require the expected score to be 4·5 on the throw in question. Then, if no other constraint is specified, the principle of maximum entropy prescribes the distribution

$$p_j = A\theta^j \qquad (j = 1, \ldots, 6)$$

where the normalising factor A and the parameter θ are to be determined by the conditions

$$\sum_{j=1}^{6} A\theta^j = 1 \qquad \sum_{j=1}^{6} Aj\theta^j = 4{\cdot}5$$

This may be regarded as the most 'spread out' of all distributions consistent with the constraint.

In the absence of any constraint other than that the sure event should have unit expectation, the principle of maximum entropy yields the uniform distribution for a finite sample space. It would therefore appear to be open to whatever objections can be raised against the classical principle of indifference as a principle for establishing prior distributions. According to the present viewpoint, however, this is not the correct way of viewing the matter. It is more accurate to regard the uniform distribution as emerging from the principle of maximum entropy, in the absence of any irredundant constraint, as a result of an implicit prior decision that the uniform distribution would best express the state of information in that situation. Indeed, many derivations of the entropy function assume this explicitly at the outset by requiring that any adequate measure must assume its maximum value for the uniform distribution.[1] This requirement is no longer appropriate in the presence of another prior distribution. In that case, whatever distribution has been chosen as prior should be the one to emerge from an application of the principle of maximum entropy in the absence of any further constraint. If the principle is to have this consequence, it is necessary to generalise the entropy concept so as to be applicable relative to any prior distribution. A generalisation of this type is well known in the case of continuous distributions. Its need in the finite case, however, has been obscured by the fact that a sample space can generally be chosen over which the uniform distribution is an adequate expression of prior opinion.[2]

Suppose that a prior distribution P^0 is given. Then the required

[1] Khinchin [1957], pp. 9–13.

[2] Even so, ambiguities can arise. Suppose that probabilities are to be assigned to the $n+1$ possible proportions of successes in a sequence of n binary trials, subject to a constraint on the expected proportion. Clearly the principle, as stated, will give a different result depending on whether it is applied to the space of $n+1$ proportions or the space of 2^n sequences.

generalisation is given by the expression

$$-\sum_{j=1}^{n} p_j \log(p_j/p_j^0)$$

where p_j^0 is the prior probability of the jth event. This reduces to the previous expression, to within an additive constant, in the case of a uniform prior. But since the new expression is never positive, it is more meaningful to define instead the *information in P relative to* P^0 as

$$I(P, P^0) = \sum_{j=1}^{n} p_j \log(p_j/p_j^0).$$

Every term for which $p_j = 0$ is understood to vanish whereas it is convenient to set $I(P, P^0) = +\infty$ if $p_j^0 = 0$ in any remaining term for which p_j is non-zero.[1]

The following property is fundamental:

$$I(P, P^0) \geqslant 0 \textit{ with equality iff } P = P^0.$$

Proof. The result does not depend on the base of the logarithms which may therefore be assumed to be natural for simplicity. If any p_j^0 vanishes whilst the corresponding p_j does not, the expression diverges and is certainly strictly positive. Suppose, therefore, that $p_j^0 = 0$ implies $p_j = 0$ for all $= 1, \ldots, n$. Since, for any positive real number x,

$$x \log x - x + 1 \geqslant 0 \textit{ with equality iff } x = 1,$$

we obtain

$$I(P, P^0) = \sum_j p_j^0 \left\{ \frac{p_j}{p_j^0} \log \frac{p_j}{p_j^0} - \frac{p_j}{p_j^0} + 1 \right\} \geqslant 0$$

where the summation need extend only over those j for which $p_j \neq 0$ and for which, consequently, $p_j^0 \neq 0$. It follows that, for these j, $I(P, P^0) = 0$ implies $p_j = p_j^0$. For the remaining j for which $p_j^0 = 0$, we have already

[1] Expressions of this type, or their extensions to the continuous case, are already studied in Good [1950], Kullback and Leibler [1951], Savage [1954] and Lindley [1956]. For more such references see Kullback [1959], ch. 1. Mention of earlier work by A. M. Turing is made in Good [1950]. A derivation of the expression $I(P, P^0)$ as the unique measure of relative information satisfying suitably modified forms of the Shannon–Khinchin conditions may be found in Hobson [1971], Appendix A. It is possible, however, to have reservations concerning such derivations on the grounds that they require that entropy, or information, should be, in a certain sense, additive. From the point of view of maximum entropy inference, however, this goes beyond the intuitions which the adequacy conditions are intended to express. For these concern themselves only with the ordering of distributions—with the idea, for example, of one distribution deviating more or less than another from the prior distribution. Additivity is, no doubt, a convenient property but it is not required for maximum entropy inference. Any strictly monotonic function of the usual expression would serve equally well. It would seem preferable, therefore, if an argument for the unique entropic ordering of distributions could be found which restricted itself to adequacy conditions formulated in terms of order alone.

seen that necessarily $p_j = p_j^0$, otherwise $I(P, P^0)$ would diverge. This completes the proof.

It follows from this that, in the absence of any constraint other than that of correct normalisation, the information in P relative to P^0 is uniquely minimised by the prior distribution as was to be required. Any other distribution contains positive information. In the case of a constraint specifying the expected score on the throw of a die, for example, the new minimum information distribution is given by

$$p_j = A p_j^0 \theta^j \qquad (j = 1, \ldots, 6)$$

where A and θ are to be determined as before, though now taking into account the prior probabilities. As in the Bayesian case, the posterior probability is found by multiplying the prior probability by a numerical factor.

The generalised principle of maximum entropy or, better, the *principle of minimum information* can now be formulated as follows:

Principle of Minimum Information: *Given the prior distribution P^0, the probability distribution P appropriate to a new state of information is one that minimises $I(P, P^0)$ subject to whatever constraints the new information imposes.*[1]

It is important to emphasise that the principle of minimum information, in this form, is not a principle for setting up prior distributions. It is, rather, a general principle of probability dynamics. It seeks to answer the question how to *modify* a probability distribution, in the light of new information, in the most conservative way. It is inapplicable in the absence of a prior distribution.[2]

As a special case, suppose that a prior distribution P^0 is given and that new evidence establishes some proposition E in the domain of P^0 with certainty. Then the new distribution incorporating this information must attribute unit probability to this event. It follows from the principle of minimum information that, provided $P^0(E)$ is non-zero, the new dis-

[1] A formulation of this principle, taking explicit account of the prior distribution, already occurs in Good [1963] as the 'Principle of Minimal Discriminability'. Good states that, according to his interpretation, its purpose is to generate null hypotheses concerning physical probability distributions which are to be tested by experiment. This contrasts, in form of expression at least, with the present interpretation according to which the purpose of the principle is to assist in the rational modification of beliefs.

[2] According to the present interpretation, the probabilities emerging from the principle of minimum information are not conditional probabilities associated with the prior distribution but unconditional probabilities of a new and entirely different distribution, unrelated to the prior distribution by the normal 'synchronic' probability calculus. This is to be understood even in the case corresponding to Bayesian conditionalisation. If this is accepted, objections of the type raised by Friedman and Shimony [1971] are not applicable.

tribution is given, for any F, by

$$P(F) = \frac{P^0(EF)}{P^0(E)} = P^0_E(F).$$

The most meaningful way of establishing this is to verify first that, for any distribution for which $P(E) = 1$, the information in P relative to P^0 is given by

$$I(P, P^0) = I(P_E, P^0_E) - \log P^0(E)$$

This is minimised when $P_E = P^0_E$ with the increase in information given by the remaining constant term $-\log P^0(E)$. The posterior distribution is thus uniquely determined given that $P(E) = 1$. Thus the Bayesian rule of conditionalisation is a special case of the principle of minimum information.[1]

It has been suggested by R. C. Jeffrey [1965], however, that the rule of conditionalisation is, strictly speaking, limited in its applicability inasmuch as it requires new information to establish some proposition with certainty. In practice there is always room for doubt. Suppose therefore that new evidence establishes some proposition E in the domain of the prior distribution P^0 only with probability q, where q may be supposed close to unity, though this is not essential. Jeffrey has proposed that the posterior distribution be given by

$$P(F) = qP^0(F|E) + \tilde{q}P^0(F|\tilde{E}) \qquad (\tilde{q} = 1-q).$$

This is the solution to a special case of the general question raised by Jeffrey as follows[2]: 'Given that a passage of experience has led the agent to change his degrees of belief in certain propositions $E_1, E_2, \ldots, E_m$ from their original values

$$P^0(E_1), P^0(E_2), \ldots, P^0(E_m)$$

to new values,

$$P(E_1), P(E_2), \ldots, P(E_m)$$

how should these changes be propagated over the rest of the structure of his beliefs?'

This is the sort of question which the principle of minimum information is designed to answer. In the special case just considered, where one takes

[1] It must be admitted, however, that this has only been demonstrated in the case where $P^0(E)$ is strictly positive whereas the Bayesian rule, to replace $P^0(F)$ by $P^0(F|E)$ when E has become certain, is asserted even when $P^0(E) = 0$, assuming the conditional probabilities to be defined independently. When $P^0(E)$ vanishes, the information in any distribution assigning positive probability to E necessarily diverges and no choice can be made on this basis. This is as it should be. Relative information has been defined only for unconditional distributions, which say nothing about the relative probabilities of events of probability zero. To deal adequately with this case it would be necessary to define the relative information of conditional distributions or, at least, the corresponding ordering.

[2] Jeffrey [1965], p. 157 adapted to the present notation

account of only a single event or proposition at a time, the principle of minimum information yields the solution advocated by Jeffrey. The same is true in the case of several mutually exclusive events. To prove this the case of a single event, it is routine to verify that provided $0 < q < 1$ we have, for any distribution P for which $P(E) = q$,

$$I(P, P^0) = qI(P_E, P^0_E) + \tilde{q}I(P^0_{\tilde{E}}, P^0_{\tilde{E}}) + \text{constant}$$

where again

$$P_E(F) = \frac{P(EF)}{P(E)} \quad etc.$$

Clearly this is minimised by separately minimising the functions on the right. Thus we obtain Jeffrey's result $P = qP^0_E + \tilde{q}P^0_{\tilde{E}}$. It is straightforward to extend this to any number of events provided they are mutually exclusive. This is in fact the most general case treated by Jeffrey. In the case where the events are not necessarily mutually exclusive, the principle of minimum information gives new results. Let $\Omega = \{\omega_1, \ldots, \omega_n\}$ be the space of possibilities. Suppose the events $E_1, \ldots, E_m \subseteq \Omega$ are constrained to have new probabilities $q_1, \ldots, q_m$ and, for convenience, write E_0 for the sure event with probability $q_0 = 1$. Assume, for simplicity, that P^0 is strictly positive and that there exists a strictly positive distribution satisfying the constraints. This means, in particular, that none of the new probabilities is 0 or 1. Then, defining the array $\{\chi_{ij}\}$ by

$$\chi_{ij} = \begin{cases} 1 & \text{if } \omega_j \in E_i \\ 0 & \text{otherwise} \end{cases} \qquad \begin{matrix} (i = 0, \ldots, m) \\ (j = 1, \ldots, n) \end{matrix}$$

the solution is given by

$$p_j = p^0_j \exp\left\{\sum_{i=0}^{m} \lambda_i \chi_{ij}\right\} \qquad (j = 1, \ldots, n)$$

with the parameters $\lambda_0, \ldots, \lambda_m$ determined by the constraints

$$q_i = \sum_{j=1}^{n} \chi_{ij} p_j \qquad (i = 0, \ldots, m).$$

There is an interesting way in which the transformation of probability distributions by the principle of minimum information appears to differ, in general, from the special case of Bayesian conditionalisation. For, in the latter case, it makes no difference whether one conditionalises successively on E_1 and E_2, in either order, or directly on the joint event E_1E_2. On the face of it, this is not true of the general case.[1] If two constraints are applied successively, there is no reason, in general, why the first should remain satisfied when information is minimised subject

[1] *Cf.* Jeffrey [1965], pp. 162–3.

to the second. The reason why it is unnecessary to maintain the first constraint explicitly in the Bayesian case is that no subsequent application of the principle of minimum information can ever reverse the decision to assign probability 1 to a given event (unless only one distribution satisfies the later constraint when no further principle is needed anyway). If then one regards the Bayesian case as one in which earlier constraints are implicitly maintained, the position is not so different in the general case. For suppose that, in the situation just treated, we define a sequence of probability distributions $\{P^i\}$ by requiring that for each $i = 1, \ldots, m$

$$P^i \textit{ minimises } I(P, P^{i-1}) \textit{ subject to } P(E_h) = q_h \textit{ for all } h = 1, \ldots, i.$$

At each stage all earlier constraints are explicitly maintained. Then it is not difficult to show that we arrive finally at the same distribution as we found before by applying all the constraints at once, which also means that the order is unimportant. This is not trivial since finally we have reached the same probability distribution by applying the same constraint to two different priors, namely P^{m-1} and P^0. In this respect the principle of minimum information behaves as well as one could expect. We shall return to this question, in greater generality, later.

It is time now to deal with the question of the effectiveness of the principle, that is with the existence and uniqueness of information-minimising distributions. The general situation is this. A prior distribution is assumed but, for one reason or another, it may no longer be an adequate expression of opinions. The new distribution should belong to a certain subset $\mathscr{C}$ of the set $\mathscr{P}$ of possible distributions. In the finite-dimensional case $\mathscr{C}$, like $\mathscr{P}$, is a subset of $\mathbf{R}^n$. $\mathscr{C}$ will normally contain many members and a choice must be made between them. The principle of minimum information prescribes the distribution in $\mathscr{C}$ that minimises $I(P, P^0)$. In principle, $\mathscr{C}$ could be an arbitrary subset of $\mathscr{P}$ and there is no guarantee that such a distribution exists. Let us see what can go wrong.

(i) $\mathscr{C}$ is empty: the constraints are inconsistent.

(ii) $\mathscr{C}$ is non-empty but $I(P,P^0)$ is infinite for all P in $\mathscr{C}$. This will occur, in the finite-dimensional case, if and only if $\mathscr{C}$ requires that a positive probability be assigned to an event of prior probability zero.

(iii) Now $\mathscr{C}$ contains distributions with finite information relative to P^0 but no minimal element. For every distribution in $\mathscr{C}$ there is another with strictly smaller information. This will arise if it is required that $P(E) > q$, where $0 < P^0(E) \leqslant q < 1$, or that $I(P, P^0) > \alpha$ for suitable $\alpha \geqslant 0$. (If $\alpha = 0$, this only says that P should be different from P^0.)

(iv) $\mathscr{C}$ contains minimal elements, with finite information relative to P^0,

but more than one. *Example*: except in certain trivial cases, the constraint that $I(P,P^0) = \alpha$ for some finite $\alpha > 0$.

We shall examine these possibilities in turn beginning with the last.

Case (iv) will never arise if $\mathscr{C}$ is *convex*. For, with P^0 fixed, $I(P, P^0)$ is a *strictly* convex function of P. Thus if P^1 and P^2 are two distinct distributions with the same finite information relative to P^0, any proper convex combination

$$\lambda P^1 + (1-\lambda)P^2 \qquad (0 < \lambda < 1)$$

has strictly smaller information. The case of convex constraints appears to be the rule rather than the exception. Certainly constraints on expectation values are of this type, however many random quantities they concern and whether they are expressed by linear equalities or weak or strong inequalities.

Case (iii) will never arise if $\mathscr{C}$ is a *closed* subset of $\mathscr{P}$ (usual topology). So far as constraints on expectation values are concerned, this means that, in general, strong (strict) inequalities are to be avoided. The fact that the principle fails to make a definite recommendation in such cases is surely a merit not a defect. Consider the simplest case where $P^0 = (\frac{1}{2}, \frac{1}{2})$, the uniform distribution over two possibilities, heads and tails say. If the constraint demands only that $P(H) > \frac{1}{2}$, it is hard to see how any principle could reasonably be expected to yield a prescription.

As was already observed, case (ii) arises if subsequent evidence requires that a strictly positive probability be assigned to an event of prior probability zero. Then it will be impossible to distinguish between distributions on the basis of their relative information as it has been defined. To obtain definite results here would require a more delicate method of comparison. An extension of the definition in this direction would certainly be of interest, though beyond the scope of this paper.[1] On the other hand it is relevant to observe that the principle itself has no inherent tendency to lead to more concentrated distributions. On the contrary, the principle is intrinsically conservative. In the case of a convex constraint, for instance, the supports of the prior and posterior distributions will coincide unless the constraint explicitly requires it to be otherwise, that is unless there is no distribution with the same support as the prior that satisfies the constraint. To this extent the principle itself contributes nothing towards the difficulties arising under (ii).

Lastly, there is nothing to be done in case (i) except to re-examine the constraints to see where one went wrong. This is the province of the 'synchronic' probability calculus.

[1] *Cf.* p. 135, n. 1.

In summary, the principle of minimum information yields a unique prescription for all closed convex constraints satisfied by at least one distribution having finite information relative to the given prior.

It is worth returning briefly to the question of the significance of the order in which constraints are applied. It was mentioned before that, in the case of Bayesian conditionalisation, it makes no difference whether one conditionalises successively on two events E_1 and E_2, in either order, or directly on the joint event E_1E_2. It was observed that a corresponding result holds in the more general case contemplated by Jeffrey provided that, at each stage, all earlier constraints are explicitly maintained. This amounts to dealing with a sequence of successively stronger constraints. The result claimed in that case was that the outcome of applying the constraints one by one is the same as applying them all at once, namely just the last. In fact it seems reasonable to expect the order to be unimportant only in such a situation. Suppose, therefore, that we have a sequence of successively stronger constraints where, indeed, without loss of generality we may suppose there to be just two: $\mathscr{C}_1 \supseteq \mathscr{C}_2$. Let P^0 be given and suppose that P^1 alone, amongst distributions in $\mathscr{C}_1$, has minimal information relative to P^0. Suppose again that P^2 alone, amongst distributions in $\mathscr{C}_2$, has minimal information relative to P^1. The question at issue is whether P^2 is the same distribution as would have been reached by applying the constraint $\mathscr{C}_2$ directly to P^0. In fact a sufficient condition for this to happen is that $\mathscr{C}_1$ should be an *affine* constraint. Then, provided $\mathscr{C}_2$ is also affine, the result can be extended to a third constraint, and so on.

An affine constraint is one that contains every probability distribution on a line through any two of its members. That is to say, along with any P_0 and P_1, it contains

$$P_\lambda = (1-\lambda)P_0 + \lambda P_1$$

for every real number λ for which P_λ remains non-negative. Constraints on expectation values are of this type. Thus if $\bar{x}_1, \ldots, \bar{x}_m$ are any real numbers and $X_1, \ldots, X_m$ are any real-valued functions on $\Omega = \{\omega_1, \ldots, \omega_n\}$, the set of probability distributions $P = (p_1, \ldots, p_n)$ such that for each $i = 1, \ldots, m$

$$\sum_{j=1}^{n} p_j X_i(\omega_j) = \bar{x}_i$$

forms a (possibly empty) affine constraint. Indeed, as is well known, this is the most general example in finite dimensions.[1]

Suppose then that $\mathscr{C}$ is an affine constraint containing at least one

[1] Previously we were dealing with the special case of an affine constraint determined by $\{0,1\}$-valued random quantities.

distribution having finite information relative to P^0. Then, necessarily, there is a unique distribution, P^1 say, in $\mathscr{C}$ having minimal information relative to P^0. Furthermore, for any distribution P in $\mathscr{C}$,

$$I(P, P^0) = I(P, P^1) + I(P^1, P^0).$$

Thus, the amounts of information in any distribution belonging to $\mathscr{C}$, relative to P^0 and P^1, differ by the same constant amount, provided they are finite. Otherwise they diverge simultaneously. This means that the functions $I(-, P^0)$ and $I(-, P^1)$ induce the same orderings in $\mathscr{C}$ and, *a fortiori*, in any stronger constraint that might be applied subsequently.

It is not possible to extend this result beyond the class of affine constraints in a straightforward way. Certainly it fails for convex constraints in general. Whether or not this should be considered a defect of the principle will depend on what alternatives exist. Nevertheless, it seems reasonable to demand of any acceptable alternative that it should possess the property in question for the restricted, but important, class of affine constraints at least. It is not obvious, however, that any principle, other than that of minimum information, meets even this limited demand. Except in the trivial[1] case $n \leqslant 3$, it is certainly not met by minimising (the square of) the Euclidean distance from the prior:

$$\sum_{j=1}^{n} (p_j - p_j^0)^2.$$

On the other hand, that principle might already be excluded on the grounds that it fails to yield the Bayesian rule of conditionalisation and has no natural infinite-dimensional generalisation. An improvement, in the latter respect at least, is afforded by the principle prescribing a distribution that minimises the distance from the prior in the sense of the metric[2]:

$$D(P, P^0) = \sup \{|P(E) - P^0(E)| : E \subseteq \Omega\}.$$

The Bayesian rule and Jeffrey's generalisation are at least optimal from this point of view, though by no means uniquely so even in elementary cases. In respect of its treatment of a sequence of successively stronger affine constraints, on the other hand, this principle fares no better in general (in cases where it succeeds in prescribing a unique posterior) than the principle based on the Euclidean metric. It is perhaps premature to

[1] If $n \leqslant 3$, any principle whatever has the property in question for a decreasing sequence of affine constraints, provided only that it preserves the prior whenever the prior already satisfies the constraint.

[2] This is essentially the principle discussed in the paper of S. May and W. Harper [1976] which was kindly brought to my attention by one of the referees. May and Harper mention the principle of minimum information as a possible 'minimum change' principle, but choose instead the principle based on a version of the supremum metric for detailed study.

conjecture that the principle of minimum information is unique in possessing the property in question. It is nonetheless possible that, by studying such transformational characteristics of various principles, a more satisfactory explanation of the peculiar reasonableness of the principle of minimum information will be found.

Appendix

Suppose that the events to which probabilities are to be assigned can be represented by a field $\mathscr{F}$ of subsets of a non-empty set Ω of arbitrary cardinality. A probability distribution over $\mathscr{F}$ is a non-negative, finitely additive[1] real-valued function on $\mathscr{F}$ such that $P(\Omega) = 1$. Let $\mathscr{P}$ be the set of all probability distributions over $\mathscr{F}$ and let P and P^0 be two such distributions. For any finite partition $\mathscr{E} = \{E_1, \ldots, E_n\}$ of Ω into sets belonging to $\mathscr{F}$, the information in P relative to P^0 with respect to $\mathscr{E}$ can be defined by

$$I_{\mathscr{E}}(P,P^0) = \sum_{j=1}^{n} P(E_j) \log \frac{P(E_j)}{P^0(E_j)}$$

which is to be understood in the same way as before when any of the terms in the summands vanish.

When confronted with a variety of finite schemes, as in the infinite case, it is only reasonable, if the aim is to minimise information, to assume no less than the worst case, though nothing positively worse. Thus it is natural to define the information in P relative to P^0 as the supremum over finite partitions:

$$I(P, P^0) = \sup \{I_{\mathscr{E}}(P, P^0)\}.$$

It is worth remarking that $I_{\mathscr{E}}(P, P^0)$ can only increase as the partition is further refined.

In this way $I(P, P^0)$ is defined for every P in $\mathscr{P}$ provided we admit the value $+\infty$. Clearly $-\infty$ is not possible value. In fact it is clear that

$$I(P, P^0) \geqslant 0 \textit{ with equality if and only if } P = P^0.$$

The distributions P for which $I(P, P^0)$ assumes a finite value will be said to belong to the *effective domain* of the function $I(-, P^0)$. A necessary, though generally not sufficient, condition for $I(P, P^0)$ to be finite is that P should be *absolutely continuous* with respect to P^0. That is to say, for

[1] The present approach endorses the viewpoint of de Finetti (see [1974], ch. 6 for example) that it is better not to assume at the outset that all probability distributions are countably additive, but to assume only those properties that follow from the meaning of probability itself—which do not include countable additivity—and to introduce the latter as a special assumption only when particular circumstances justify.

any $\epsilon > 0$ there should exist a $\delta > 0$ such that, for every E in $\mathscr{F}$,

$$P^0(E) < \delta \text{ implies } P(E) < \epsilon.$$

In this case, if P and P^0 are countably additive over a σ-field, a Radon–Nikodym derivative dP/dP^0 exists and the equivalent integral expression

$$I(P, P^0) = \int_\Omega \log(dP/dP^0)\, dP$$

can be obtained (Guiaşu [1977], ch. 2).

In general, without making the assumption of countable additivity, it can be shown that

$$I(-, P^0) \text{ is strictly convex over its effective domain}$$

for any prior distribution P^0. Furthermore, if we endow $\mathscr{P}$ with the relativised product topology by considering $\mathscr{P}$ as a subset of the product $\mathbf{R}^{\mathscr{F}}$, it follows that

$$I(-, P^0) \text{ is lower semi-continuous}$$

for any prior distribution P^0.

From the last result, together with the well-known compactness of $\mathscr{P}$, it follows that $I(-, P^0)$ assumes its minimum value on any closed subset $\mathscr{C}$ of $\mathscr{P}$. That is, there exists a distribution P^1 in $\mathscr{C}$ such that

$$I(P^1, P^0) = \inf \{I(P, P^0) \colon P \in \mathscr{C}\}.$$

Furthermore, if $\mathscr{C}$ is convex and includes a distribution having finite information relative to P^0, strict convexity of $I(-, P^0)$ implies that P^1 is unique. Thus, in general, the principle of minimum information yields a unique prescription for all closed convex constraints satisfied by at least one distribution having finite information relative to the given prior.

It was observed in the finite-dimensional case that the principle has no inherent tendency to lead to either more or less concentrated distributions. In general this is expressed by the mutual absolute continuity of prior and posterior distributions. Let us assume that P^1 has minimal finite information over $\mathscr{C}$ relative to P^0. Then certainly P^1 is absolutely continuous with respect to P^0 since $I(P^1, P^0)$ is finite. Conversely, if $\mathscr{C}$ is convex, P^0 is absolutely continuous with respect to P^1 unless there is no distribution in $\mathscr{C}$ having finite information relative to P^0 and with respect to which P^0 is absolutely continuous. Recalling that no purely finitely additive distribution is absolutely continuous with respect to a countably additive distribution, it follows that from a countably additive prior the principle can only lead to a countably additive posterior.[1] On the other hand, the principle leads from a purely finitely additive prior to a

[1] By a purely finitely additive distribution is meant one that is finitely but not countably additive.

countably additive posterior only if there is no purely finitely additive distribution in $\mathscr{C}$ (assumed to be convex) having finite information relative to P^0. So that although countably additive distributions are 'absorbing' in this sense, the principle displays no inherent attraction towards them.

It is of interest to note that the extension of distributions by means of countable additivity has no effect on their relative information. If P, P^0 are countably additive distributions over the field $\mathscr{F}$ and $\hat{P}$, $\hat{P}^0$ are their unique countably additive extensions to the σ-field generated by $\mathscr{F}$, then

$$I(\hat{P}, \hat{P}^0) = I(P, P^0)$$

where it is to be understood that the two amounts of information are calculated by means of finite partitions of the extended and restricted fields, respectively.

All that was said before in the finite case concerning Bayesian conditionalisation as an example of the principle of minimum information holds good in general. The same is true of Jeffrey's rule when there are finitely many events whose posterior probabilities are prescribed. In the general case, however, Jeffrey's problem may be formulated with respect to the prescription of infinitely many posterior probabilities. Provided some distribution having finite information relative to the given prior satisfies this prescription, the principle of minimum information again leads to a unique solution.

The question of the significance of the order in which constraints are applied requires more delicate and extended discussion in the general case. This will be dealt with elsewhere.

The University of Sussex

REFERENCES

DE FINETTI, B. [1974]: *Theory of Probability*. London: John Wiley & Sons.

FRIEDMAN, K. and SHIMONY, A. [1971]: 'Jaynes's Maximum Entropy Prescription and Probability Theory', *J. Statist. Phys.* **4**, pp. 381–4.

GOOD, I. J. [1950]: *Probability and the Weighing of Evidence*. London: Charles Griffin & Co. Ltd.

GOOD, I. J. [1963]: 'Maximum Entropy for Hypothesis Formulation, especially for Multi-dimensional Contingency Tables', *Ann. Mat. Statist.* **34**, pp. 911–34.

GUIAŞU, S. [1977]: *Information Theory with Applications*. New York: McGraw-Hill, Inc.

HOBSON, A. [1971]: *Concepts in Statistical Mechanics*. New York: Gordon and Breach, Inc.

JAYNES, E. T. [1957]: 'Information Theory and Statistical Mechanics', *Phys. Rev.* **106**, pp. 620–30; **108**, pp. 171–90.

JAYNES, E. T. [1968]: 'Prior Probabilities', *IEEE Trans. Syst. Science and Cybernetics*, SSC-4, pp. 227–41.

JEFFREY, R. C. [1965]: *The Logic of Decision*. New York: McGraw-Hill, Inc.

KHINCHIN, A. I. [1957]: *Mathematical Foundations of Information Theory*. New York: Dover Publications, Inc.

KULLBACK, S. and LEIBLER, R. A. [1951]: 'On Information and Sufficiency', *Ann. Math. Statist.* **22**, pp. 79–86.

KULLBACK, S. [1959]: *Information Theory and Statistics*, New York: John Wiley & Sons, Inc.

LINDLEY, D. V. [1956]: 'On a Measure of the Information Provided by an Experiment', *Ann. Math. Statist.* **27**, pp. 986–1005.

MAY, S. and HARPER, W. [1976]: 'Towards an Optimisation Procedure for Applying Minimum Change Principles in Probability Kinematics'. In: W. L. Harper and C. A. Hooker (*eds.*), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, vol. III, pp. 137–66. Dordrecht: D. Reidel Publishing Company.

SAVAGE, L. J. [1954]: *The Foundations of Statistics*. New York: John Wiley & Sons, Inc.