Pramāṇa – J. Phys., Vol. 26, No. 4, April 1986, pp. 301–310. © Printed in India.

On the axiomatic approach to the maximum entropy principle of inference

S N KARBELKAR

Joint Astronomy Program, Physics Department, Indian Institute of Science, Bangalore 560012, India

MS received 17 August 1985; revised 18 January 1986

Abstract. Recent axiomatic derivations of the maximum entropy principle from consistency conditions are critically examined. We show that proper application of consistency conditions alone allows a wider class of functionals, essentially of the form $\int dx\, p(x)\,[p(x)/g(x)]^s$, for some real number s, to be used for inductive inference and the commonly used form $-\int dx\, p(x) \ln [p(x)/g(x)]$ is only a particular case. The role of the prior density $g(x)$ is clarified. It is possible to regard it as a geometric factor, describing the coordinate system used and it does not represent information of the same kind as obtained by measurements on the system in the form of expectation values.

Keywords. Inductive inference; maximum entropy principle; prior distribution.

PACS No. 02·50; 03·65; 05·20

1. Problem of inductive inference

In many practical situations measured expectation values of functions of state of a system (even when noise free) are insufficient to uniquely determine the underlying probability distribution. For example, in radio astronomy (Bracewell and Roberts 1954) one is faced with a problem of restoring the sky brightness distribution from partial knowledge of its Fourier coefficients which can be measured as correlations using the aperture synthesis technique. The problem of inductive inference is to choose, amongst all possible distributions consistent with the data, a distribution which is 'the best' in some sense. One, then, starts with a functional of the probability distribution which is considered to be a measure of its 'goodness'. The 'best' distribution is one for which, subject to constraints of measured data, the functional attains its global maximum. Note that 'the best' distribution depends on what functional is chosen as a quantitative measure of 'goodness'.

Based on Shannon's information theory, Jaynes (1957) proposed the maximum entropy principle (MEP) which uses the thermodynamic entropy, which is also the Shannon's information measure,

$$-\sum_i p_i \ln p_i,$$

as a functional to be used for inductive inference for the following reason. Statistical mechanical description of a system contains far less information than is contained in

mechanical description; the thermodynamic entropy is a measure of missing information due to availability of only macroscopic measurements. Likewise, to every probability distribution one can assign an information measure which expresses, quantitatively, the uncertainty associated with statistical description of the system. Jaynes advocated Shannon's measure not because the expression is same as thermodynamic entropy but because he expected that any other information measure may eventually lead to contradiction since no other information measure is additive for independent sources of uncertainty. Recently there have been attempts (Shore and Johnson 1980, 1983; Tikochinsky *et al* 1984) to show that the thermodynamic entropy functional has the deeper significance, attached to it by Jaynes, as the only consistent procedure of statistical inference when insufficient number of expectation values are known. These axiomatic derivations do not appeal to subjective characterization of information measure but instead impose certain consistency requirements on the methods of inference. From this point of view, pioneered by Jaynes, statistical mechanics is to be considered as a special case of inductive inference, not dependent on auxiliary hypothesis of equal apriori probabilities, ergodicity etc for its validity.

In § 2 a brief summary of these axiomatic derivations of the MEP is given. In § 3 it is pointed out that the application of consistency conditions needs some care. Consistency conditions, when properly defined, do not restrict the functional to the thermodynamic entropy. In § 4 an interpretation of the prior density, which is crucial to application of a variational principle (even if form of the functional is derived), is given.

2. Summary of axiomatic derivations of the MEP

Axiomatic derivations by Shore and Johnson (SJ) and by Tikochinsky *et al* (TTL) are based on a common fundamental principle: Two ways of using the same information should lead to the same result.

In the case of continuous probability densities it is necessary to invoke, apart from the probability density to be induced from the measurements and called the posterior (or equivalently inferred density), a density called the prior which, according to SJ, is an apriori estimate of the unknown true system density. We discuss the need for the prior and its interpretation in § 4.

Two 'equivalent' ways of using the same information appear in the derivation by SJ when one considers two systems with states labelled x_1 and x_2 and given information is separable in x_1 and x_2 i.e., no information about interaction between the two systems is available. In this case one way is to solve the problem of inductive inference separately for the two systems and obtain separate posteriors $p_1(x_1)$ and $p_2(x_2)$ from the given priors $g_1(x_1)$ and $g_2(x_2)$ respectively. When systems are independent one can construct a joint posterior $p_{12}(x_1, x_2) = p_1(x_1)p_2(x_2)$. Note that it is necessary to assume the unmeasured correlation. The second way is to treat both the systems together in terms of joint densities. Now it is necessary to choose the joint prior density $g_{12}(x_1, x_2)$ since its marginals $g_1(x_1)$ and $g_2(x_2)$ do not determine its form uniquely. SJ make the choice $g_{12}(x_1, x_2) = g_1(x_1)g_2(x_2)$. Once a prior is given (or chosen) the posterior is uniquely determined for a particular variational principle. So let the posterior $P_{12}(x_1, x_2)$ be obtained from the chosen prior $g_{12}(x_1, x_2)$ and using the sum total information. In both these ways it was necessary to make some statistical hypothesis in order to get a unique posterior. These hypothesis were to assume (i) that the joint posterior

factorizes, in the first way and (ii) that the joint prior factorizes, in the second way. SJ, in their system independence axiom, take these two ways to be equivalent and therefore require the posteriors obtained in these two ways to be equal i.e.,

$$p_{12}(x_1, x_2) = P_{12}(x_1, x_2).$$

Apart from the system independence axiom, stated above, SJ also require, as a part of consistency condition uniqueness, invariance under coordinate transformations and subset independence axiom which allows one to treat an independent subset of the system in terms of a conditional density.

TTL formulate their consistency conditions for the case of a reproducible experiment, where it is possible to carry out independent repetitions (which need not be large in number) of an experiment. They consider the following two ways of giving information equivalent:

(a) The result of the basic experiment is $m+1$ expectation values

$$\langle A_r \rangle = \sum_{i=1}^{n} A_{ri} p_i \qquad r = 0, 1, \ldots, m < n \tag{1}$$

$$A_{0i} \equiv 1$$

of state variables A_r which take values A_{ri} when the system is in state i. Here p_i is the (unknown) true probability associated with the state i; there are n mutually exclusive and exhaustive states.

(b) Let the basic experiment be repeated N times. One can, then, define sample averages

$$B_{r\tilde{N}} = \frac{1}{N} \sum_{i=1}^{n} N_i A_{ri} \qquad r = 0, 1, \ldots, m$$

$$\sum_{i=1}^{n} N_i = N$$

for a particular realization $\tilde{N} \equiv (N_1, \ldots, N_n)$ in which the state i occurred N_i times (order immaterial). One can specify averages $\langle B_r \rangle$ as a result of the N-repetition experiment:

$$\langle B_r \rangle = \sum_{\tilde{N}} B_{r\tilde{N}} P_{\tilde{N}} \tag{2}$$

$$P_{\tilde{N}} = \frac{N!}{N_1! \ldots N_n!} p_1^{N_1} \ldots p_n^{N_n},$$

where $P_{\tilde{N}}$ denotes the probability of occurrence of a particular realization $\tilde{N}$ when repetitions are independent. When repetitions are independent specifying the information in (2) is the same as specifying the information in (1).

The inductive inference algorithm (denoted by TTL as) A induces a set of n probabilities, say, $q_1, \ldots, q_n$ from the information in (1) and a set of $l = {}^{N+n-1}C_{n-1}$ probabilities, say, $Q_{\tilde{N}}$ from the information in (2). TTL require the algorithm A to be uniform in the sense that it acts on the data of a given kind in the same way i.e., when

inducing probabilities $Q_{\tilde{N}}$'s it attaches no special significance to the symbol n (this is very reasonable as a realization $\tilde{N}$ can be thought of as a state of some superexperiment). The algorithm A is said to be consistent if the induced probabilities $Q_{\tilde{N}}$'s are related to the induced probabilities q_i's in the same way the true probabilities $P_{\tilde{N}}$'s (for independent repetitions) are related to the p_i's of the basic experiment.

Both these derivations by SJ and TTL conclude that MEP is the only consistent method of inference.

3. Criticism of the axiomatic derivations of the MEP

Stated for independent repetitions, consistency condition due to TTL is reasonable. However, the flaw in their derivation is that the apriori information that the repetitions are independent is not used explicitly in the form of constraints. The need to use these constraints can be seen from the following argument.

Consider a modified experiment which is N-repetition of the original experiment; however, the repetitions are not independent. Let the probability of occurrence of a particular realization $\tilde{N}$ be $P'_{\tilde{N}}$. In this case, one may still get the same $m+1$ expectation values

$$\langle B_r \rangle = \sum_{\tilde{N}} B_{r\tilde{N}} P'_{\tilde{N}} \qquad r = 0, 1, \ldots, m \tag{3}$$

as before, though, now $P'_{\tilde{N}} \neq P_{\tilde{N}}$ as the repetitions are not independent. The algorithm A can be used to induce, in a 'uniform' way a set of l probabilities, say $Q'_{\tilde{N}}$ from the information in (3). It is, however, incorrect to require these $Q'_{\tilde{N}}$'s to be related to the q's of the basic experiment in any simple way. In fact, when the repetitions are correlated the $P'_{\tilde{N}}$'s of the N-correlated repetition experiment cannot be related to the p_i's of the basic experiment, the probabilities p_{ij}'s of occurrence of state i in jth repetition must be specified. Therefore, in general, specifying expectation values for an experiment and for an experiment which is N-repetition of that experiment are not equivalent ways of taking 'the same' information into account. Only in the case that it is known apriori that repetitions are independent, one is justified in imposing consistency conditions. The data (for example (2) or (3) which is numerically the same but a result of different experiment), by itself, does not tell us whether the repetitions are independent or otherwise. *That the repetitions are independent is a significant apriori knowledge* and constraints corresponding to this knowledge *must* be used in addition to the $m+1$ constraints given in (2). These constraints are not used by TTL and the variational principle is required to do this job.

It is shown in Appendix A that when these extra constraints are used the algorithm is not restricted to one using $-\sum p_i \ln p_i$ for 'entropy', but a wider class of functionals, essentially of the form $\sum p_i^s$ for some real number s, is allowed.

Now, when it is not known, one may model the N-repetition experiment as N independent repetitions of the original experiment. This, however, is a part of statistical modelling and does not constitute consistency conditions, for one could, if the situation demands, model the repetitions as correlated ones. Thus, in the absence of knowledge of a certain correlation in the problem at hand, one should keep modelling the correlation apart from consistency conditions (if any); one should not give the status of

a consistency axiom to any statistical modelling. This is what the system independence axiom by SJ does, as is discussed below.

As seen before, in the case of separable information about two systems, it was assumed either that the joint posterior factorizes or that the joint prior factorizes; only then it was possible to get a posterior. Now assuming a factorized posterior is assuming independence of the systems, and it remains to be seen if factorizable prior is equivalent (in the case of separable information) to assuming factorizable posterior. SJ interpret (Shore and Johnson 1980, p. 28) the prior in a data-adaptive manner in that a prior is our apriori estimate of the true system density. A factorized prior, according to SJ, represents our apriori estimate that the systems are independent. Now the new information is separable and thus gives no clue to any interaction between the systems and therefore there is no need to change our apriori knowledge about the independence of the systems. This is the justification given by SJ of their system independence axiom which treats factorizable posterior and factorizable prior (in the case of separable measurements) on the same footing. It is shown in § 4 that a prior cannot be interpreted in a data-adaptive manner in the sense that statistical information (such as system independence) cannot be encoded on the prior. A factorizable prior (whether the information is separable or otherwise) is not equivalent to a factorizable posterior and the two ways of getting a posterior, considered by SJ, are not equivalent. These two ways are equivalent if and only if it is known apriori that the systems are independent. However, one must use constraints corresponding to this in the second way of obtaining the joint posterior. Since these constraints are not used by SJ their system independence axiom always models independent systems whenever prior factorizes and separable information is available. Note that the notion of 'separable information' is coordinate frame-dependent. It is necessary that the prior factorizes in the same frame in which information is separable in order that the problem of inductive inference factorizes into two problems, one each for the two systems. We call a separable information (expectation values involve either x_1 or x_2 but not both x_1 and x_2) to be separable in the geometric sense if the prior factorizes in the same frame; and separable in the statistical sense if the systems are statistically independent in that frame. Taking an information separable in the geometric sense to mean statistically separable information is putting unmeasured correlation equal to zero. This is analogous to obtaining a 'dirty map' (or principal solution) in radio astronomy where unmeasured Fourier coefficients of the sky brightness distribution are put equal to zero (Bracewell and Roberts 1954).

It is shown in Appendix B that when constraints corresponding to system independence are used a wider class of functionals, of the form $\int \mathrm{d}x\, p(x)\, [p(x)/g(x)]^s$, is allowed.

4. Interpretation of the prior density

In the case of continuous densities, the only variational principle which is invariant under coordinate transformations and uses the posterior (to be induced) alone is the one that uses functionals equivalent to $\int \mathrm{d}x\, p(x)$. We see that such a functional is of no use as a principle of inductive inference. It is, therefore, necessary to invoke yet another density, say $g(x)$, so that the variational principle has sufficient richness of its solutions. At this stage the prior appears for purely mathematical reason and it is worthwhile to interpret it.

First we show that a prior cannot be interpreted in a data-adaptive manner. By data-adaptive we mean the following. Let us start with a prior, say $g_1(x)$, and obtain posterior $p_1(x)$ using new information in the form of expectation values. Then, one cannot drop these constraints and use the posterior $p_1(x)$ as prior when some more information is available. The posterior $p_2(x)$ obtained in this way will, in general, be different from the posterior $p(x)$ obtained from $g_1(x)$ using both sets of constraints.

To see this, we consider the entropy $-\int dx\, p(x) \ln [p(x)/g(x)]$, though the argument is quite general. Let

$$\int dx\, p_{\text{true}}(x) A_k(x) = \Omega_k \qquad k = 1, \ldots, m_1,$$

$$\int dx\, p_{\text{true}}(x) B_j(x) = \Gamma_j \qquad j = 1, \ldots, m_2,$$

be two sets of expectation values, given to us. We assume that such expectation values are realizable in that there exist p's which satisfy these. The posterior $p(x)$ obtained from $g_1(x)$ and using both these sets of constraints is of the form

$$p(x) = g_1(x) \exp\left\{\sum_{k=1}^{m_1} \Lambda_k A_k(x) + \sum_{j=1}^{m_2} \alpha_j B_j(x)\right\},$$

where the Lagrange multipliers Λ's and α's are chosen (uniquely) such that $p(x)$ satisfies the given expectation values. The posterior $p_1(x)$ obtained from $g_1(x)$ using only the first set of constraints is of the form

$$p_1(x) = g_1(x) \exp\left\{\sum_{k=1}^{m_1} \lambda_k A_k(x)\right\},$$

where the Lagrange multipliers λ's are chosen such that $p_1(x)$ satisfies the first set of constraints. In general $\lambda_k \neq \Lambda_k$ for some values of k as the Lagrange multipliers Λ's depend on $B_j(x)$ and Γ_j while λ's do not depend on these. Now, the posterior $p_2(x)$ obtained using $p_1(x)$ as prior and the second set of constraints is of the form

$$p_2(x) = p_1(x) \exp\left\{\sum_{j=1}^{m_2} \beta_j B_j(x)\right\}$$

$$= g_1(x) \exp\left\{\sum_{k=1}^{m_1} \lambda_k A_k(x) + \sum_{j=1}^{m_2} \beta_j B_j(x)\right\}$$

for properly chosen Lagrange multipliers β's. Since, in general, $\lambda_k \neq \Lambda_k$ and in order to satisfy $\int dx\, p_2(x) B_j(x) = \Gamma_j$ we have $\beta_j \neq \alpha_j$ for some values of j. Thus we see that $p_2(x) \neq p(x)$. The posterior $p_2(x)$ satisfies the second set of constraints, but not the first set of constraints. *A prior is not equivalent to a set of constraints and thus cannot represent our apriori knowledge in the form of expectation values.*

To see that such a density is natural to the continuous case consider a discrete system with mutually exclusive and exhaustive states labeled $1, \ldots, n$. These states can also be labeled by distinct real numbers, say, $x_1, \ldots, x_n$ where $x_1 < x_2 < \ldots < x_n$ and all x's belong to some interval, say $[a, b]$. One may ask: What is the number of system states in some interval, say, $(x - \frac{1}{2}dx,\ x + \frac{1}{2}dx]$? In the limit that $n \to \infty$ one may be able to define a density $g(x)$ such that $g(x)dx$ gives the relative number of states in the dx neighbourhood of x. $g(x)$ is a scalar density with respect to coordinate transformations, it is non-negative and need not be normalizable. Note that, so far, we have not defined

probability distribution on the system states. The scalar density $g(x)$ then tells us how the system was 'composed'. In the discrete case, considered above, a natural n-tuple was $(p_1, \ldots, p_n)$ of probabilities to be induced. By 'natural' we mean an object that enters the very statement of the problem of inductive inference. Consider a problem of inductive inference for a system which has n states as before, but now, each state has a degeneracy, say g_i. Now, the two n-tuples, natural to the problem of inductive inference, are $(p_1, \ldots, p_n)$ and $(g_1, \ldots, g_n)$. The latter n-tuple is also associated with the first system considered above where every g_i is unity. The degeneracy n-tuple is a apriori knowledge about the system in that it tells us the relative weight attached to each state. However, the statement that all g's are equal is not equivalent to the statement that our apriori estimates of the p's are all equal. For, the notion of degeneracy is independent of our estimates (or even knowledge) of the true state probabilities.

Thus, in the variational principle of the form

$$\int \mathrm{d}x\, p(x) h[p(x)/g(x)]$$

for some function h, the function $g(x)$ can be interpreted as a non-negative scalar density devoid of any statistical significance as an 'estimate'. This density is analogous to the degeneracy n-tuple of the discrete case.

5. Conclusion

The prior density, which is crucial to application of a variational-principle-approach to inductive inference is shown to be free of any data-adaptive interpretation i.e., it is shown that statistical information about the system cannot be encoded on the prior and thereby it cannot represent our apriori estimate of the unknown true system density. Prior is interpreted as a non-negative scalar density (with respect to coordinate transformations) analogous to degeneracy of a discrete system.

Derivation of functionals, to be used in the variational principle, will, in general, involve consistency requirements, formulated in terms of existence of more than one equivalent ways of using the same information, as the very general requirements of uniqueness and coordinate invariance do not fully restrict the form of the functional. These equivalent ways will be defined for a specific experiment and it is necessary to remember this in the form of appropriate constraints. When such constraints are used in the case of specific experiments considered by Shore and Johnson and by Tikochinsky *et al* it is shown that the MEP is not the only consistent method of inductive inference, functionals of the form $\int \mathrm{d}x\, p(x) [p(x)/g(x)]^s$, for some real number s, give consistent results. Jaynes conjectured that because of nonadditivity, functionals other than $-\Sigma p_i \ln p_i$ will eventually lead to a contradiction. We have shown that functionals, of the above form, exist which have the nonadditivity property and yet do not lead to any inconsistencies in the kind of experiment under consideration. Consistency conditions alone do not lead to the inference procedure of $-\Sigma p_i \ln p_i$ maximization; some more input is necessary. The Jaynes conjecture remains to be proved.

Acknowledgements

It is author's pleasure to thank Dr Rajaram Nityananda for introducing him to the field and for numerous instructive discussions. He is also grateful to Dipankar Bhattacharya

for a perusal of the manuscript. Thanks are due to Raman Research Institute for research facilities.

Appendix A

Consistent functionals in the case of reproducible experiments

We show that the 'entropy'

$$H[p; g] = \sum_{i=1}^{n} p_i (p_i/g_i)^s$$

for some real s, such that H is a convex functional of p, is self-consistent in the case of N independent repetitions. In the case of the basic experiment $g_i = 1$ and $p_i^{(1)}$'s are given by

$$(1+s)(p_i^{(1)})^s + \sum_{k=0}^{m} \lambda_k^{(1)} A_{ki} = 0$$

where the Lagrange multipliers are such that $\sum_{i=1}^{n} A_{ri} p_i^{(1)} = \langle A_r \rangle$. The superscript (j) refers to j-repetition experiment. The constraint that the N repetitions are independent is most easily effected by writing

$$P_{\hat{N}} = g_{\hat{N}} (p_1^{(N)})^{N_1} \ldots (p_n^{(N)})^{N_n}; \quad \sum_{i=1}^{n} p_i^{(N)} = 1$$

$$g_{\hat{N}} = \frac{N!}{N_1! \ldots N_n!},$$

and so

$$H[P_{\hat{N}}; g_{\hat{N}}] = \left[\sum_{i=1}^{n} (p_i^{(N)})^{1+s} \right]^N.$$

Information obtained for N-repetition experiment is

$$\langle B_r \rangle = \sum_{i=1}^{n} A_{ri} p_i^{(N)},$$

therefore $p_i^{(N)}$'s are given by

$$\left[\sum_{i=1}^{n} (p_i^{(N)})^{1+s} \right]^{N-1} (1+s)(p_i^{(N)})^s + \sum_{k=0}^{m} \lambda_k^{(N)} A_{kj} = 0.$$

That the solution is the same as before i.e. $p_j^{(N)} = p_j^{(1)}$ can be seen from the choice

$$\lambda_k^{(N)} = \lambda_k^{(1)} \left[\sum_{i=1}^{n} (p_i^{(1)})^{1+s} \right]^{N-1}.$$

Note that this is possible because of the appearance of an overall factor

$$\left[\sum_{i=1}^{n} (p_i^{(N)})^{1+s}\right]^{N-1},$$

which can be absorbed in the Lagrange multipliers. This may not be the case for more general 'entropies'. For example, a sum or difference of two distinct power laws may lead to inconsistencies i.e. $p_i^{(N)} \neq p_i^{(1)}$ even after the constraint that repetitions are independent is explicitly used as above. It can also be shown that 'entropies' $-\Sigma g_i \ln[p_i/g_i]$ and $-\Sigma p_i \ln[p_i/g_i]$ are consistent.

Appendix B

Consistent functionals in the case of independent systems

The constraint that the two systems labeled x_1 and x_2 are independent is most easily effected by writing $p_{12}(x_1, x_2) = p_{1J}(x_1)p_{2J}(x_2)$ subject to two constraints $\int dx_1\, p_{1J} = \int dx_2\, p_{2J} = 1$. The subscript J represents the fact that the problem of inductive inference is being solved for the joint system (the 'second' way in the text). Consider an entropy of the form

$$H[P_{12}; g_{12}] = \iint dx_1\, dx_2\, g_1(x_1) g_2(x_2) F\left[\frac{p_{1J}(x_1)p_{2J}(x_2)}{g_1(x_1)g_2(x_2)}\right].$$

For $p_{1J}(x_1)$ to equal the solution $p_1(x_1)$ obtained for the system 1 alone (the 'first' way in the text) it is sufficient and (perhaps) necessary that the variation $\delta H/\delta p_{1J}$ equal $F'(p_{1J}/g_1)$ apart from a multiplicative and an additive factor which can be a functional of p_{2J} alone i.e.,

$$\int dx_2\, p_{2J} F'\left(\frac{p_{1J}p_{2J}}{g_1 g_2}\right) = F'\left(\frac{p_{1J}}{g_1}\right)\int dx_2\, g_2 G\left(\frac{p_{2J}}{g_2}\right) + \int dx_2\, g_2 R\left(\frac{p_{2J}}{g_2}\right)$$

for some functions G and R for all p_{2J}. Here the prime denotes derivative with respect to the argument. Functional derivative of this equation with respect to p_{2J} gives

$$r_{1J} r_{2J} F''(r_{1J} r_{2J}) + F'(r_{1J} r_{2J}) = G'(r_{2J}) F'(r_{1J}) + R'(r_{2J}); \quad r = p/g.$$

We also require the solution $p_{2J}(x_2)$ to equal $p_2(x_2)$. This amounts to the requirement that the functions G' and R' equal F' apart from a multiplicative and an additive constant i.e.

$$r_{1J} r_{2J} F''(r_{1J} r_{2J}) + F'(r_{1J} r_{2J}) = \lambda F'(r_{1J}) F'(r_{2J}) + \alpha(F'(r_{1J}) + F'(r_{2J})) \quad (B1)$$

for some λ and α. To determine the function F we differentiate this equation with respect to r_{1J} and then let $r_{1J} = x, r_{2J} = 1$ to get the homogeneous equation

$$xF'''(x) + F''(x)(2 - \alpha - \lambda F'(1)) = 0,$$

whose solution is of the form $F(x) = x^m$ where

$$m(m-1)(m-\mu) = 0; \quad \mu = \alpha + \lambda F'(1).$$

The roots of this indicial equation are all distinct if $\mu \neq 0$ and $\mu \neq 1$ and the function F is of the form

$$F(x) = a_1 + a_2 x + a_3 x^{\mu},$$

where the constants a's are chosen to satisfy the equation (B1). The form of the function F when $\mu = 0$ or $\mu = 1$ is

$$F(x) = b_1 + b_2 \ln x + b_3 x \qquad \mu = 0,$$
$$= c_1 + c_2 x + c_3 x \ln x \qquad \mu = 1,$$

where the constants b's and c's are chosen to satisfy the equation (B1). Thus we see that the 'entropies' of the form

$$\int dx\, p(x)\,[p(x)/g(x)]^s \qquad \text{(and the limiting forms}$$
$$-\int dx\, g(x) \ln [p(x)/g(x)];\ -\int dx\, p(x) \ln [p(x)/g(x)]\text{)}$$

are consistent when a separable information is given for statistically independent systems.

References

Bracewell R N and Roberts J A 1954 *Aust. J. Phys.* 7 615
Jaynes E T 1957 *Phys. Rev.* **106** 620; **108** 171
Shore J E and Johnson R 1980 *IEEE Trans. Inf. Theory* **26** 26 (SJ)
Shore J E and Johnson R 1983 *IEEE Trans. Inf. Theory* **29** 943
Tikochinsky Y, Tishby N Z and Levine R D 1984 *Phys. Rev. Lett.* **52** 1357 (TTL)