homepagethanksemail

Entropic Inference for Assigning Probabilities: Some Difficulties in Axiomatics and Applications

Alberto Solana-Ortega and Vicente Solana

Instituto de Matemáticas y Física Fundamental, CSIC
Serrano 123, Madrid 28006, Spain

Abstract. The importance of entropic notions for assigning probabilities is highlighted. We review the main interpretations and uses of entropy functionals and methods. We also examine the justifications offered to support them, in particular the attempts to axiomatically derive a unique expression for entropic procedures in compliance with rationality and consistency requirements of a canon of plausible inference. The main difficulties arising when trying to apply these methods are pointed out. Ultimately they manifest the incompleteness of inference theory.

Keywords: <Missing keywords>
PACS: <Missing classification>

INTRODUCTION

The question[1234567891011121314151617181919] how to assign probabilities is inescapable in order to develop a quantitative theory of plausible inference for reasoning about the plausibilities or certainty degrees of conjectures, on the only basis of partial evidential knowledge insufficient to determine their truth. Yet this numerical assessment is an extraordinarily difficult and controversial problem, especially when available evidences are limited and observational data scarce. Modernly, it was conceived as a methodological issue, pertaining to the application of probability theory, and hence left aside from theoretical considerations. Compared to axiomatic studies of probability, focused on the relation between individual inferences and the derivation of rules to combine probabilities, formal approaches for obtaining procedures to solve such inferences have received little attention. However, along the past century it has regained acceptance the idea that assigning probabilities is not a matter of special purpose techniques of dubious soundness, but constitutes a fundamental epistemological problem, which should be supported on rational grounds [41]. Besides a greater emphasis in logic, a novel perspective has come up, namely that plausible inference is holistic and refers to all possible conjectures in a language, so the assignment of probabilities corresponds to the selection of distributions.

Among the most attractive procedures for this task we find entropic methods, characterized by the selection of the distribution which extremizes an entropy functional subject to a set of constraints representing a particular kind of evidential knowledge, more specifically high-order information about probabilities, also called testable information, which usually takes the form of restrictions setting values to distribution moments.

Several milestones can be pointed out in the way to an epistemic conception of entropic inference [40]. First, the introduction of technical notions to globally describe and quantify the state of knowledge in a situation of uncertainty regarding the conjectures, viz. Fisher's information in statistics (1925); Shannon [37] and Wiener's information-theoretic entropies, with precursors in statistical physics and communications analysis; Good's [15] expected weight of evidence, preceded by the investigations towards an extended logic of Peirce (1878), Keynes [32] and Jeffreys [27], and by the cryptographic applications of Turing during World War II; the semantic definition of information and entropy of Carnap and Bar-Hillel [6][5]; and the generalized distances between probability distributions devised by Mahalanobis (1936), Bhattacharya (1943), Jeffreys (1946)

---

and Kullback [34], which are the basis for information gain measures.

Secondly, the acknowledgement of the complementarity of these entropic notions with probability. This is specially clear in the works of Good, Ingarden, Urbanik and Domotor, who, starting from an analysis of the information concept alone, without previous recourse to probabilities, established a structural link between probabilities and entropy, allowing to define the latter in terms of the former [15][21][11]. From another perspective, Cox speculated that the coupling of these concepts indicates a duality between the logic of assertions and a logic of questions [8] (see [33]).

And in the third place, recognition by Jaynes [22], Ingarden [20], Good [16] and Kullback [34] that this complementarity of entropy and probability can be employed to construct procedures to assign unknown probabilities starting from the processing of known information. Notwithstanding the practical approaches of Gibbs and Shannon, the first explicit proposal of a general scientific inference rule based on entropic notions was due to Jaynes, who enunciated the Principle of Maximum Entropy (MaxEnt), prescribing the selection of the distribution which maximizes Shannon's absolute entropy

$$H_{Shannon}(Q) = -\sum_{i=1}^{n} q_i \log q_i, \tag{1}$$

subject to linear constraints on the distribution $Q = (q_1, ..., q_n)$.

Independently, Kullback proposed a rule for inference when a distribution $P = (p_1, ..., p_n)$ is considered in addition to the constraints. This rule, known as the Principle of Minimum Cross Entropy (MinxEnt), picks as solution the distribution which minimizes, taking into account the constraints representing the available testable knowledge, the Kullback-Leibler (KL) relative entropy functional

$$D_{KL}(Q \parallel P) = \sum_{i=1}^{n} q_i \log \frac{q_i}{p_i}. \tag{2}$$

Later on, generalized formulations have been advocated [1][36][29][9], stipulating the extremization of extended entropies, as follows:

$$\tilde{Q} = \arg\min_{\mathcal{E}_q} D_{gen}(Q \parallel P), \tag{3}$$

where $\mathcal{E}_q$ denotes a feasible space of distributions Q satisfying a set of linear constraints, and $D_{gen}(Q \parallel P)$ a generalized relative entropy functional. Common choices are

$$D_{Csiszár}(Q \parallel P) = \sum_{i=1}^{n} p_i \phi\left(\frac{q_i}{p_i}\right), \tag{4}$$

with ϕ convex and twice differentiable, such that $\phi(1) = 0$, which comprise the family

$$D^{\alpha}_{Rényi}(Q \parallel P) = \frac{1}{\alpha - 1} \log \sum_{i=1}^{n} q_i^{\alpha} p_i^{1-\alpha}, with\, \alpha > 0, \alpha \neq 1, \tag{5}$$

of additive Rényi entropies, the KL measure corresponding to the limiting case $\alpha \to 1$.

INTERPRETATION OF ENTROPY MEASURES AND METHODS

Entropy functionals have been interpreted in many ways, for instance as measures of disorder, randomnesss, surprise, freedom of choice, diversity and distinguishability. Here we will be concerned only with their intepretation in inference theory.

Shannon entropy (1) measures the amount of global uncertainty expressed by a probability distribution representing a state of knowledge about the conjectures in a language system, given an evidence. In other words, it quantifies the missing information which remains to be learned to reach certainty, that is, the expected information that an eventual confirmation of one conjecture would provide. It may also be regarded as a measure of uniformity, since a distribution representing greater uncertainty is more spread out.

Several interpretations have been put forward for KL entropy (2). Originally, Good proposed it as a measure of the expected weight of evidence provided by a new possible observational evidence x, in favour of a hypothesis H_1 implying a probability distribution Q, against another hypothesis H_2 implying P. On the other hand, Kullback regarded it as a natural directed divergence from a probability distribution Q to a fixed distribution P adopted as an estimate, and interpreted it as the mean information per possible sample of observations for discriminating in favour of a hypothesis H_1 implying Q against a hypothesis H_2 implying P, when H_1 is assumed to be true. Thus both authors were implicitly considering the functional (2) as $D_{KL}(Q(x \mid H_1) \parallel P(x \mid H_2))$, having as arguments distributions conditioned on hypothesis H_1 and H_2, respectively. (Note that no reference is made in the former interpretations to an updating operation from P to Q.)

Next, Rényi [36] introduced the idea that relative entropy measured the expected information gain to be obtained if distribution Q is used instead of P. But only with Hobson [19] it started to be viewed as an information gain associated with the passing from an initial P to a final Q. In particular, when constraints on Q are taken into account, KL entropy can be adopted as quantifying the information provided by these constraints to update P into Q, an interpretation which has become a standard after [28][38].

The previous interpretations have led to various uses of entropic procedures. First, the MaxEnt method, characterized by the maximization of Shannon entropy (1), has been regarded as a procedure for direct assignment of probabilities [29][14], by selecting, once the constraints expressing the relevant testable information are taken into account, the distribution $\tilde{Q}$ which maximizes our global uncertainty about conjectures. More exactly, Jaynes [26] defended its use within the Bayesian probability framework for choosing prior distributions for parameters, as a generalization of the classical principle of insufficient reason when high-order information is at hand.

Kullback and Good [34][16] utilized MinxEnt for three purposes: the indirect assignment of probabilities $\tilde{Q}$ from the knowledge of a default distribution P and a set of constraints; the classification of samples of observational data to the closest population from a family of populations represented by candidate distributions P; and the testing of hypothesis using the KL functional as a statistic. Note that the latter two uses pertain to direct assignment problems. However, following [38], the main use nowadays of MinxEnt is as a learning rule associated with a knowledge revision problem in which an initial model P is updated, by the addition of constraints, into an optimal distribution $\tilde{Q}$, the closest to P while consistent with the information constraints provide, geometrically interpretable [10][9] as the non-linear projection of P upon the feasible space of distri-

butions $\mathcal{E}_q$ specified by the testable information. This knowledge revision view provides an alternative to Bayes probability updating rule when new information is incorporated in the form of constraints, so the question arises which one should be used. (See [25][44] for an examination of the different positions on their validity and mutual relations.)

In spite of conceding the possibility of interpreting MinxEnt as a probability updating rule [25], Jaynes [26] understood entropic methods essentially as direct probability assignment procedures, with a marked objective character. Thus, he only considered in the discrete case the Shannon absolute entropy. In the continuous case, invariance arguments led him to extend this absolute entropy to the entropic functional

$$S[Q,m] = -\int q(\theta) \log \frac{q(\theta)}{m(\theta)} d\theta, \tag{6}$$

where $m(\theta)$, the preprior, is a measure of the limiting density of discrete elementary disjunct conjectures in the domain of parameter θ. Although the mathematical form of (6) is, except for the negative sign, identical to the extension of KL functional (2) to the continuum, its interpretation is very different, for, although $m(\theta)$ could play the role of a probability density function, it need not to, and could be regarded just as a degeneracy factor associated with the domain of θ. The difficulty with (6), nevertheless, is that $m(\theta)$ is in general not given, so the problem of assigning probabilities and selecting an optimal Q is eluded, by transforming it into another one, concerning the choice of $m(\theta)$.

JUSTIFICATION OF ENTROPIC INFERENCE METHODS

Contrary to other techniques, which in spite of being recognizably 'ad hoc' are uncritically used, there was from the beginning the aspiration of establishing entropic inference methods on solid grounds, even more intense after the attacks for apparently allowing getting probabilities 'ex nihilo', for being representation dependent under refinement of languages, and for conflicting with Bayesian updating. In view of the multitude of entropies and methods that can be formulated, one may ask whether there is a "best" functional or a privileged procedure to make inference. Jaynes defended that, when the available information takes the form of linear constraints on probabilities, the preferred method should be MaxEnt. Many different reasons support this preference, e.g. pragmatic success, combinatorial arguments, information theory and logical consistency. Next we examine how definitive are these arguments.

Regarding *pragmatism*, the success of entropic procedures is indubitable in many areas, starting with the derivation of the classical results of statistical physics and communications theory, and the solution of inverse problems in geophysics, econometrics and image analysis. Simplicity of formalism and computational efficiency are its main appeals. Tractability and practical advantages, however, are not fundamental criteria, since they depend on the developmental stage of optimization techniques, where major improvements may always take place, and are difficult to formalize. On the other hand, there are fields where different functionals allow to select in a more direct way distributions of theoretical significance, which need not belong to the family of exponential distributions associated with MaxEnt, and may give better results [29]. Finally, in critical disciplines, such as risk analysis of extreme events, the idea that a method is good

because "it works" does not apply, since observations are usually scarce and cannot be used to compare with predictions. Hence pragmatism is insufficient.

Another kind of justification supporting Shannon and KL entropies is based on *combinatorics* and asymptotics. The original argument was put forward by Boltzmann, who in the multinomial case defined entropy as a multiplicity of a distribution of particles in different microstates and proposed the selection of the distribution of maximum multiplicity, which consequently can be "realized" in the greatest of ways. After a suggestion of Wallis, Jaynes developed this approach in [24]. He also investigated how far lie other distributions admitted by the constraints with respect to the selected maximum entropy distribution, proving the entropy concentration theorem, which states that for a large number of experiments, the fraction of distributions satisfying the constraints outside an arbitrary neighbourhood is exponentially small, so that the majority of feasible distributions are concentrated close to the MaxEnt distribution. (See [17] for a critique regarding the incorporation of constraints in the limiting process.) More general results for relative entropy had already been offered by Kullback [34]. Recently another relevant property was demonstrated, namely the convergence in the asymptotic limit of Bayesian conditioning and MinxEnt [3]. Both results have been unified in a unique strong concentration theorem [18]. But analysis of these large deviations arguments [9] shows that their applicability is restricted to situations with very large numbers of observations, typically statistical physics problems. They do not cover the majority of applications, in particular those pertaining to the modelling of extraordinary phenomena, where little evidence is available, so cannot be regarded as universal grounds for a general foundation.

Inititally, the *formal* justifications for Shannon entropy were based on its properties as a measure of uncertainty and information. Besides his original axiomatic derivation, relying on the mathematical expression of information in terms of probabilities, and its successive improvements by Khinchin (1957), Fadeev (1958) and Lee (1964), other axiomatic approaches were proposed which made no reference to probabilities [15][21]. From a different viewpoint, Carnap and Bar-Hillel obtained the same expression for semantic information [6]. Analogous axiomatic derivations of weight of evidence, relative entropy, information gain and directed divergence were also presented [15] [19] [28]. However, it soon became clear that the notions of information and uncertainty are too rich to be represented by only one functional, and many generalized expressions for entropy were obtained from alternative sets of axioms [6][36][1]. Therefore, these approaches cannot attribute absolute correction to Shannon and KL functionals.

Despite their importance, the former justifications have only provided partial support, lending high plausibility to MaxEnt and MinxEnt methods, but failing to set a unitary foundation for all sorts of applications. The most promising approach for this purpose is the *consistency-based* axiomatic derivation of entropy procedures within the logical probability framework of [32][27], analogous to the derivations of probability calculation rules of Cox [7] and Carnap [4], in conformity with the canon of plausible inference, a set of elementary requirements of rationality and morality so essential that every scientific inference theory should satisfy to avoid contradictions [41].

Jaynes conjectured that MaxEnt was the only logically consistent solution [22]. To support his claim, he defended that Shannon's axiomatics were really consistency arguments. Now since the latter refer only to the obtention of an entropy functional for measuring uncertainty, and not to its extremization subject to constraints, he had to supply

them with a reasoning based on further application of the canon. When all the relevant information in a situation is taken into account, the maximum reduction in uncertainty is achieved. The remaining ambiguity to determine a distribution should then be resolved by applying a policy of honesty, frankly acknowledging the extent of our ignorance by considering every possibility allowed by the available information. Specifically, probabilities should be selected in such a way as to maximize, with respect to what is unknown, the global uncertainty in a given state of knowledge. The resulting distribution will be then the most rational, least biased, least prejudiced and less committed with what is not given. Another choice would amount to presupposing more information than really available, violating information conservation requirements, more exactly the information expliciteness and consistency demands proposed in [41]. The difficulty with this reasoning is twofold. First [43], not all axioms express consistency properties. For example, the importance of the additive property for independent events seems to respond more to calculation reasons than to common sense [29]. In addition, the axioms are valid only for discrete problems, so the uncertainty interpretation does not extend easily to the continuum. Moreover, consistency refers to the functional, not to the inference procedure. In any case, MaxEnt is not the only method which allows to recover constraints without adding or eliminating information. It is not even the only entropic method based on a reasonable representation of uncertainty. Hence its identification with a principle of honesty is too vague to characterize just one inference procedure.

The landmark in the foundations of entropic inference was due to Shore and Johnson [38], who made a decisive turn when they considered that the justification of entropy methods should be based on the properties characterizing the notion of inference procedure. Their two core ideas were: a) To represent inference, when the starting point is the high-order evidential knowledge expressed by linear constraints I on probability distributions, as a logical *operator*, either as a direct assignment rule

$$\tilde{Q} = \mathcal{F}[I] \tag{7}$$

for inferring a feasible distribution $\tilde{Q}$ consistent with I, or as an updating rule

$$\tilde{Q} = \mathcal{G}[P, I] \tag{8}$$

when a reference distribution P is given too. And b) to derive the form of the logical operator by imposing compliance with the canon of plausible inference, without referring to any subjective explanation of information. In particular they emphasized the self-consistency requirement stipulating that reasonable inference methods should lead to the same results when following admissible alternative ways of taking into account equivalent information. Four axioms satisfying this requirement were then proposed: i) uniqueness of the inferred distribution $\tilde{Q}$; ii) invariance under the choice of coordinates; iii) system independence, that is invariance under decomposition of a joint system when only independent information about its *independent* components is available; and iv) subset independence, namely invariance under conditioning of a system on its disjoint subsets. From these axioms they derived a set of functional equations whose solution is unique, and hence considered proven that MinxEnt and MaxEnt are the only consistent methods when the available information corresponds to expected value constraints, the resulting form of the functionals to be extremized being just a corollary.

Motivated by this approach, similar derivations have been presented. Skilling [39] extended it for inferring arbitrary positive functions, and also addressed the selection of reference default models. Paris and Kern-Isberner [35][31] have investigated in more depth the logical conception of inference in relation to common sense and conditionals, paying attention to the representation and revision of knowledge bases using formal languages. In contrast to previous derivations, they reached the same results without assuming 'ab initio' that the operators $\mathcal{F}$ and $\mathcal{G}$ correspond to a variational problem. Csiszár [10][9] has examined what alternative axioms would have to be adopted in order to obtain different procedures, such as generalized entropic methods and other statistical techniques. Specifically, he derived MinxEnt and MaxEnt independently from two combinations of axioms. However, other generalized entropy methods, based for instance on the extremization of (4), result when starting from different groupings.

A distinct consistency approach was advanced in [42] for the discrete case. Only two conditions, of consistency under *independent* repetitions of an experiment, and of uniformity in the way constraints are taken into account, were required, and no variational representation in terms of extremization of a functional was assumed. In [13] claim was also made that MaxEnt is the only consistent rule for direct assignment of probabilities.

DIFFICULTIES IN AXIOMATICS AND APPLICATIONS

What, then, is the state of the art of the rational foundations of entropic procedures? Where do we stand in relation to their applicability for inference purposes? Apart from minor criticisms, the logical consistency approach of Shore and Johnson, and the sequel of Tikoshinsky et al., were received very favourably. Nevertheless, Karbelkar [30] and Uffink [43] independently examined their original arguments and found obscure points and errors. Specifically, the critical issue concerns the incorporation of the idea of *independence* into the axioms and proofs for the main theorems.

In [38], the axioms are phrased twice, first in an informal manner, and then formally. The informal statement demands equivalence between the description of a joint system in terms of a joint inferred probability distribution and the representation of its *independent* components in terms of marginal inferred distributions, when the available information refers separately to each subsystem and imposes no joint restriction on the inferred distribution. But the formal axiom and the proof of uniqueness of MinxEnt omit this condition of independence for the components, and require a much stronger property, namely equivalence whether or not the system components are in fact independent. It is well known that, under separability of the constraints expressing testable information and factorization of the joint reference distribution P into its marginals, MinxEnt implies the factorization of the joint inferred distribution $\tilde{Q}$. This has been interpreted as a special principle of insufficient reason with regard to statistical dependence, since in absence of reason for correlations between components, the method selects a distribution which does not consider any. However, the former entailment does not mean that separability of testable information pertaining to each of the components and factorization of P imply MinxEnt, and thus the factorization of $\tilde{Q}$. Karbelkar and Uffink convincingly argue that the requirement of logical consistency cannot be applied in this case, because

the two alternative ways of modelling the joint system really start from non-equivalent information, so the proof in [38], which pretends such an equivalence, is incorrect. In addition, they demonstrate that the explicit inclusion of the condition of system independence and factorization of $\tilde{Q}$, which is what ensures the equivalence of the informations adopted as starting points, leads as consistent solution to Rényi's entropies (5).

In [30][43] it is also discussed that the derivation in [42] is flawed, because it requires an assumption of independence between repetitions of an experiment that is not explicitly incorporated in the mathematical representation of prior knowledge. When this presupposition is taken into account, the family of Rényi entropies results as well. We note also that a similar derivation [12], applied to the so-called Judy Benjamin problem, allows for the generalized entropies of Rényi. In this respect we recall that the axiomatics of Csiszár lead to various entropic methods too, depending on the groups of axioms assumed. In [10] the difficulty was stressed of setting preferences among axiom systems, and hence of ranking procedures. However, in this case the reasoning is typical of modern mathematics, for it stops at axioms level, when, as underlined in [41], no progress is possible without the wider viewpoint that axiomatizations are to be justified on the basis of an extramathematical canon of compelling common sense demands.

In sum, the existence of solutions to the problem of determining consistent inference procedures has been proven, but not their uniqueness. We know that MaxEnt and MinxEnt, which still make in our opinion the best candidates, do not violate the demands of the inference canon, for they are solutions in all consistency derivations. Nonetheless, until uniqueness is not dilucidated, the foundations of inference will be incomplete.

This theoretical incompleteness entails in its turn a practical difficulty, namely the selection of a form for an entropy procedure in real applications, which adds to the other main difficulties encountered when putting into use the entropy formalisms: the election of a default reference probability distribution and the choice of the constraints expressing high-order testable information.

Regarding the *selection of reference distributions* P in the quasi-ignorance situation when no testable information is at hand, consideration of one default model or another is decisive, since the inferred distribution $\tilde{Q}$ may change radically depending on our choice. Following [27], Jaynes [23][26] proposed the use of invariance arguments to select these functions. He defended that the symmetries involved in the physical conceptualization of parameters should be taken into account, by choosing references which are invariant under the symmetry group of transformations. This proposal is nonetheless subject to discussion [2], because no general constructive procedure was established, and its application to multiparametric inferences in the continuum, where symmetries are not self-evident, poses difficulties and may lead to distributions with undesirable properties.

Finally, concerning the *choice of constraints* [24][26][30][29][43][44][17], the majority of statistical inference methods assume as starting point an initial information expressed probabilistically, for example by means of constraints or through the likelihood function. Usually it is accepted that this information represents the available empirical evidence. The difficulty is due to the strong dependence of inferences on the exact expression of this kind of evidential knowledge, which moreover can be too complicated and artificial, with no clear physical meaning. Entropic methods constitute no exception to this circumstance, although they are not to be criticized more than others, since the problem is universal. In practical terms the question is how to select among representa-

tions supposedly expressing basic evidential knowledge, especially observational data, in view of the enormous diversity of possibilities to define moments and other restrictions on probabilities. The challenge, however, is to justify such a selection and respond to why should basic evidential knowledge be represented in terms of probabilities or expected value constraints, when these evidences don't originally have this form. Or, more generally, to solve the old scientific question of how to connect empirical data to probability assignments and attribute physical content to inference. There are interpretative aspects involved that make this probabilistic representation anything but automatic.

We recall the importance of three facets here: the empirical representativity of sample functions with respect to observational data; the theoretical representativity of expected values with respect to probability distributions; and the representativity of sample functions with regard to expected values, i.e. the validity of their identification through a so-called constraint rule, or, what is the same, the admissibility of sample functions as estimators. When few observations are available, as in the case of extraordinary events, this identification is problematic and unjustified, for estimators are not reliable, yet high-order properties are required to shape the tails of probability distributions.

Therefore, new arguments are needed to determine probability representation rules. This issue, on which formalisms are silent, beyond its practical dimension ultimately manifests the need for more investigations on the logical expression and encoding of basic evidences, in order to make progress towards a unified inference theory.

ACKNOWLEDGMENTS

Research was funded through grants REN2002-011337 and VEM2006-26947E of DGI, Spain. We thank the kind support from the MaxEnt2006 Organizing Committee.

REFERENCES

1. J. Aczél and Z. Daroczy, *On Measures of Information and their Characterization*, New York: Academic Press, 1975.
2. J. M. Bernardo and A. F. M. Smith, Bayesian Theory, Chichester: John Wiley Sons, 1994.
3. J. van Campenhout and T. M. Cover, "Maximum entropy and conditional probability", *IEEE Trans. Inform. Theory* **IT-27**, 483–89 (1981).
4. R. Carnap, *Logical Foundations of Probability,* Chicago: University of Chicago Press, 1950.
5. R. Carnap, *Two Essays on Entropy*, Berkeley: University of California Press, 1977.
6. R. Carnap and Y. Bar-Hillel, "An outline of a theory of semantic information", Technical Report No. 247, Research Laboratory in Electronics, Massachussets Institute of Technology, 1952.
7. R. T. Cox, *The Algebra of Probable Inference*, Baltimore: The John Hopkins Press, 1961.
8. R. T. Cox, "Of inference and inquiry". In *The Maximum Entropy Formalism*, ed. by M. Tribus and R. D. Levine, Massachussets Institute of Technology, 1979.
9. I. Csiszár, "MaxEnt, mathematics and information theory". In *Maximum Entropy and Bayesian Methods,* edited by K. M. Hanson and R. N. Silver, Dordrecht: Kluwer, 1996, pp. 35–50.
10. I. Csiszár, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems", *Annals of Statistics* **19**, 2032-2066 (1991).
11. Z. Domotor, "Qualitative Information and Entropy Structures". In *Information and Inference*, edited by J. Hintikka and P. Suppes, Dordrecht, Reidel, 1970, pp. 148–196.
12. B. C. van Fraassen, "A problem for relative information minimizers in probability kinematics", *British Journal for the Philosophy of Science* **32**, 375–379 (1981) and **37**, 453–463 (1986).

13. A. J. M. Garrett, "Maximum entropy from the laws of probability". In *Bayesian Inference and Maximum Entropy Methods*, A. Mohammad-Djafari, Melville, NY: AIP, 2001, pp. 3–22.
14. A. Golan, G. Judge and D. Miller, *Maximum Entropy Econometrics: Robust Estimation with Limited Data,* Chichester: John Wiley Sons, 1966.
15. I. J. Good, *Probability and the weighing of evidence*, London: Charles Griffin, 1950.
16. I. J. Good, "Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables", *Annals of Mathematical Statistics* **34**, 911–934 (1963).
17. M. Grendár Jr. and M. Grendár, "What is the question MaxEnt answers?". In *Maximum Entropy and Bayesian Methods,*, edited by A. Mohammad-Djafari, Dordrecht: Kluwer, 2001, pp. 83–93.
18. P. Grünwald and A. P. David, "Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory", *Annals of Statistics* **32**, 1347–1433 (2004).
19. A. Hobson, "A new theorem of information theory", *J. Stat. Phys.***1**, 383–391 (1969).
20. R. S. Ingarden, "Information theory and variational principles in statistical theories", *Bull. Acad. Polon. Sci. Ser. Sci. Math. Astronom. Phys.* **11**, 541–547 (1963).
21. R. S. Ingarden and K. Urbanik, "Information without probability", *Colloq. Math.***9**, 131–150 (1962).
22. E. T. Jaynes, "Information theory and statistical mechanics I", *Physics Review* **108**, 171–90 (1957).
23. E. T. Jaynes, "Prior probabilities", *IEEE Trans. Syst. Sci. Cybern.* **4**, 227–241 (1968).
24. E. T. Jaynes, "On the rationale of maximum-entropy methods", *IEEE Trans. Inform. Theory* **70(9)**, 939–952 (1982).
25. E. T. Jaynes, "The relation of bayesian and maximum entropy methods". In *Maximum Entropy and Bayesian Methods*, G. J. Erickson and C. R. Smith, Dordrecht: Kluwer, 1988, pp. 267–281.
26. E. T. Jaynes, *Probability Theory: The Logic of Science*. G. L. Bretthorst (Ed.). Cambridge: Cambridge University Press, 2003.
27. H. Jeffreys, *Theory of Probability.* Oxford: The Clarendon Press, 1939, second edition 1948.
28. R. W. Johnson, "Axiomatic characterization of the directed divergences and their linear combinations", *IEEE Trans.Inform. Theory* **IT-25**, 709–716 (1979).
29. J. N. Kapur and H. K. Kesavan, *Entropy Optimization Principles with Applications,* San Diego: Academic Press, 1992.
30. S. N. Karbelkar, "On the axiomatic approach to the maximum entropy principle of inference", *Pramana-Journal of Physics* **26(4)**: 301–310 (1986).
31. G. Kern-Isberner, *Conditionals in Nonmonotonic Reasoning and Belief Revision*, Berlin: Springer, 2001.
32. J. M. Keynes, *A Treatise on Probability.* London: MacMillan and Co, 1921.
33. K. Knuth, "Lattice duality: The origin of probability and entropy", *Neurocomp.* **67C**: 245-274 (2005).
34. S. Kullback, *Information Theory and Statistics,* New York: John Wiley & Sons, 1959.
35. J. B. Paris, *The Uncertain Reasoner's Companion: A Mathematical Perspective*, Cambridge: Cambridge University Press, 1994.
36. A. Rényi, "On measures of entropy and information". In *Proceed. 4th Berkeley Symp. Math. Stat. Probability*, Vol 1, pp. 547–561, University of California Press, Berkeley, 1961.
37. C. E. Shannon, "A mathematical theory of communication", *Bell System Technical Journal* **27**, 379–423 and 623–656 (1948).
38. J. E. Shore and R. W. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy", *IEEE Trans. on Information theory* **IT–26(1)**, 26–37 (1980).
39. J. Skilling: "The axioms of maximum entropy". In *Maximum Entropy and Bayesian Methods*, edited by G. J. Erickson and C. R. Smith, Vol. 1, Dordrecht: Kluwer, 1988, pp. 173-187.
40. A. Solana-Ortega, "The information revolution is yet to come (An homage to Claude Shannon)". In *Bayesian Inference and Maximum Entropy Methods*, R. Fry, Melville, NY: AIP, 2002, pp. 458–473.
41. A. Solana-Ortega and V. Solana. "Another look at the canon of plausible inference". In *Bayesian Inference and Maximum Entropy Methods*, K. H. Knuth et al., Melville, NY: AIP, 2005, pp. 382–391.
42. Y. Tikochinsky et al., "Consistent inference of probabilities for reproducible data", *Physical Review Letters* **52**, 1357–1360 (1984).
43. J. Uffink, "Can the maximum entropy principle be explained as a consistency requirement?", *Studies in History and Philosophy of Modern Physics* **26B**, 223–261 (1995).
44. J. Uffink, "The constraint rule of the maximum entropy principle", *Studies in History and Philosophy of Modern Physics* **27**, 47–79 (1996).