Environment and Planning B, 1975, volume 2, pages 99-105

# Generalized measures of information, Bayes' likelihood ratio and Jaynes' formalism

L March
Department of Systems Design, University of Waterloo, Ontario, N2L 3GI, Canada
M Batty ¶
Department of Civil Engineering, University of Waterloo, Ontario, N2L 3GI, Canada
Received 4 July 1975

**Abstract.** This paper relates generalized measures of information to expected likelihood functions (ELFs) derived from Bayes' equation. It then demonstrates that Jaynes' formalism may be extended to formulate a *class* of minimally-prejudiced models of which those derived from Shannon's measure are but a limiting and special case. The rôle of probable inference and of information-minimizing models in design is commented on.

## Introduction

Information theory, from its first formal presentation by Shannon (1948) in the field of communications, has found extensive use elsewhere. In particular, since the mid-sixties a considerable number of applications have been described in design and planning studies. To a large extent these applications have been concerned with two problems: theory-construction and model-building. Both problems are synthetic in nature: the first takes observations from a number of situations, and from these *general* hypotheses may be inferred; the second asserts the truth of certain hypotheses and observations (or specifications), on the basis of which a *singular* situation, in accord with these assumptions, may be generated in some optimal way. The first problem is essentially scientific, the second is a design activity in Simon's (1969) general sense: it is, for example, the quintessential mode of reasoning in law and in medical diagnostics where individual 'cases' have to be proven, as well as in design as it is more traditionally understood. Neither problem is deterministic in the manner of analytic deductive reasoning. Both have solutions that, at best, can be stated in the form 'there is a probability $p_x^{oyz}$ that $x$ is true given that $y$ and $z$ are accepted as true explicitly along with other implicit facts $o$ in the case'. The former mode of reasoning is *inductive*, the latter *abductive* to use C. S. Peirce's term (see Feibleman, 1970). The distinction between these two modes of synthetic reasoning has been developed by March (1975), where an attempt at a Bayesian interpretation of the design activity is made and support is given to Tribus' (1969) theory of design. Design here is seen as an information-processing activity, the dynamics of which have been studied in an exploratory paper by Batty (1974).

In this paper certain fundamental issues of importance to an appreciation of the logic of design are addressed. The paper is presented in a questioning spirit in the hope that understanding will be enriched by having to face questions hitherto ignored. The paper also presents a unified view of the information-minimizing methods of probable inference developed directly from Bayes' theorem, accepting Cox's (1961) axiomatic treatment of the algebra up to the statement of the theorem.

¶ Address from September 1975: Department of Geography, University of Reading, Reading RG6 2AF, Berkshire, England.

**Bayes' theorem**

Bayes' theorem relates prior and posterior probabilities by means of the likelihood ratio. We may express the theorem thus:

$$p_i^{oa} = p_i^o L^i \,, \tag{1}$$

where $p_i^{oa}$ is the posterior probability given new information '$a$', $p_i^o$ is the prior probability before information '$a$' is introduced, and $L^i$ is the likelihood ratio which makes it possible to estimate $p_i^{oa}$ from $p_i^o$ using information '$a$'. Clearly,

$$L^i = \frac{p_i^{oa}}{p_i^o} \,, \tag{2}$$

is finite and positive, providing $p_i^o \neq 0$, which we shall assume from now on. $L^i$ is the likelihood ratio associated with individual probabilities $p_i^o$ and $p_i^{oa}$. For the systems represented by $\{p_i^{oa}\}$ and $\{p_i^o\}$ we may calculate the expected likelihood ratio (ELR)

$$L = \sum_i p_i^{oa} L^i \,. \tag{3}$$

Equally we could calculate expected values for monotonic functions of $L^i$ such as

$$(L^i)^{\alpha-1} = \left(\frac{p_i^{oa}}{p_i^o}\right)^{\alpha-1} \,, \qquad \alpha > 0 \,, \quad \alpha \neq 1 \,, \tag{4}$$

or,

$$\ln L^i = \ln \frac{p_i^{oa}}{p_i^o} \,. \tag{5}$$

Since $\alpha = 2$ gives equation (3) as a special case of equation (4), we shall investigate the general case for $\alpha > 0$, but $\alpha \neq 1$, in the first instance. We shall call equations (4) and (5) expected likelihood functions (ELF). The ELF in the general case, $\alpha \neq 1$, is thus given by

$$L_\alpha = \sum_i p_i^{oa} \left(\frac{p_i^{oa}}{p_i^o}\right)^{\alpha-1} = \sum_i (p_i^{oa})^\alpha (p_i^o)^{1-\alpha} \,. \tag{6}$$

It can be shown that the ELF has a stationary point when $p_i^{oa} = p_i^o$, and that its value then is unity. If $\alpha > 1$ the stationary point is a minimum, but for $\alpha < 1$ it is a maximum. We shall now transform the ELF further so that a new function ${}_1\Lambda_\alpha \geqslant 0$ is formed, having as its properties

$${}_1\Lambda_\alpha(\{p_i^{oa} = p_i^o\},\{p_i^o\}) = 0 \,, \tag{7}$$

and

$${}_1\Lambda_\alpha(\{1,0\},\{\tfrac{1}{2},\tfrac{1}{2}\}) = {}_1\Lambda_\alpha(\{0,1\},\{\tfrac{1}{2},\tfrac{1}{2}\}) = 1 \,. \tag{8}$$

These are arbitrary conditions giving a scale of values for ${}_1\Lambda_\alpha$. The transformation we consider is

$${}_1\Lambda_\alpha = AL_\alpha - B \,, \tag{9}$$

where $A$ and $B$ are constants to be determined.

From equation (7) we have

$$\begin{aligned} {}_1\Lambda_\alpha(\{p_i^{oa} = p_i^o\},\{p_i^o\}) &= A\sum_i (p_i^{oa})^\alpha (p_i^o)^{1-\alpha} - B = A\sum_i p_i^{oa} - B \\ &= A - B = 0 \,. \end{aligned} \tag{10}$$

Therefore $A = B$, and from equation (8) we then have

$$ {}_1\Lambda_\alpha(\{1,0\},\{\tfrac{1}{2},\tfrac{1}{2}\}) = A[1^\alpha(\tfrac{1}{2})^{1-\alpha} + 0^\alpha(\tfrac{1}{2})^{1-\alpha} - 1] = A(2^{\alpha-1} - 1) $$
$$ = 1 , \tag{11} $$

so that $A = (2^{\alpha-1} - 1)^{-1}$. The transformation which gives the desired properties (7) and (8) is

$$ {}_1\Lambda_\alpha = (2^{\alpha-1} - 1)^{-1}(L_\alpha - 1) . \tag{12} $$

We might have taken another tack. Since the ELF has a stationary point at the value 1 and is always positive definite, the function ${}_2\Lambda_\alpha = A \ln L_\alpha \geqslant 0$ has a stationary point at zero, so that

$$ {}_2\Lambda_\alpha(\{p_i^{oa} = p_i^o\},\{p_i^o\}) = 0 , \tag{13} $$

as in equation (7). The condition (8) requires that

$$ {}_2\Lambda_\alpha(\{1,0\},\{\tfrac{1}{2},\tfrac{1}{2}\}) = A \ln[0^\alpha(\tfrac{1}{2})^{1-\alpha} + 1^\alpha(\tfrac{1}{2})^{1-\alpha}] = A \ln 2^{\alpha-1} $$
$$ = A(\alpha - 1)\ln 2 , \tag{14} $$

whence $A = [(\alpha - 1)\ln 2]^{-1}$.

Another suitable transformation would thus be:

$$ {}_2\Lambda_\alpha = [(\alpha - 1)\ln 2]^{-1} \ln L_\alpha \tag{15} $$
$$ = (\alpha - 1)^{-1} \operatorname{lb} L_\alpha , \tag{16} $$

where lb is logarithmus to base 2.

It is an elementary exercise to confirm that ${}_1\Lambda_\alpha$ and ${}_2\Lambda_\alpha$ are always positive and have a unique minimum $\Lambda_\alpha(\{p_i^{oa} = p_i^o\},\{p_i^o\}) = 0$.

The case $\alpha = 1$ may now be examined by taking the limiting case $\alpha \to 1$ both for ${}_1\Lambda_\alpha$ and ${}_2\Lambda_\alpha$. The result, using d'Hôpital's rule, is the same in both cases:

$$ {}_1\Lambda_1 = {}_2\Lambda_1 = (\ln 2)^{-1} \sum_i p_i^{oa} \ln \frac{p_i^{oa}}{p_i^o} \tag{17} $$

$$ = \sum_i p_i^{oa} \operatorname{lb} \frac{p_i^{oa}}{p_i^o} . \tag{18} $$

The same form is derived directly from equation (5), where the expected value of $\ln L^i$ is given by

$$ \sum_i p_i^{oa} \ln \frac{p_i^{oa}}{p_i^o} $$

which differs from equation (18) only by a constant factor related to the choice of logarithmus.

**Transformations of the ELF and measures of information**

It will be clear that the last form, equation (18), is Kullback's (1959) measure of directed divergence. Although perhaps less familiar, the other forms are measures of information put forward by Renyi (1961) and Taneja (1974), both of whom point out that the limiting case $\alpha \to 1$ gives Kullback's measure. More recently Hobson and Cheng (1973) have referred to directed divergence as a measure of information which generalizes Shannon's proposal (Shannon, 1948) in the case of unequal priors, and which overcomes problems of measure (Jaynes, 1968) when the discrete summation is extended to the continuous case. Taneja's measure for equal priors, in

essence, appears earlier in the literature in papers by Havdra and Charvát (1967), and Daróczy (1970).

In view of these similarities it is proposed to rename terms as follows:

$$ {}_1\Lambda_\alpha \triangleq {}_T I_\alpha \,, \qquad \alpha \neq 1 \tag{19} $$

$$ {}_2\Lambda_\alpha \triangleq {}_R I_\alpha \,, \qquad \alpha \neq 1 \tag{20} $$

$$ {}_1\Lambda_1 = {}_2\Lambda_\alpha \triangleq {}_{KH} I \,, \qquad \alpha \to 1 \tag{21} $$

where the prefixed subscripts stand for Taneja, Renyi, and Kullback/Hobson respectively. Note that

$$ {}_{KH} I = {}_S I \,, \qquad \text{when} \quad p_i^o = \frac{1}{n} \,, \ \forall i \tag{22} $$

where the new prefix stands for Shannon.

We have thus demonstrated a direct relationship between Bayes' likelihood ratio and measures of information. Information-minimizing models (Evans, 1969)—akin to entropy-maximizing models (Jaynes, 1957; Tribus, 1969)—are thus seen to be an extension of the Bayesian approach. Indeed effectively these models are seen to seek stationary values of the ELR with respect to the posterior probabilities, under constraints embracing additional information. Such models are then understood to estimate posteriors by modifying the priors through the likelihood ratio, so that the overall divergence between the two is the least possible compatible with the new information. The models may be said to be least prejudiced.

**Jaynes' formalism extended**

An obvious generalization of Jaynes' formalism may now be stated: given
(a) a set of prior probabilities $\{p_i^o\}$, $p_i^o \neq 0$, $\forall i$,
(b) a set of expected mean values $\{\langle a^k \rangle\}$ of functions $a^k(x)$ of a discrete random variate $x$, $k = 0, 1, 2, ..., m$, with
(c) $a^o(x_i) = 1 = \langle a^o \rangle$, $\forall i$,
we may obtain an estimate for the likelihood ratio which will modify [in the least-prejudiced way with respect to the information contained in (b)] the prior $p_i^o$ to give $p_i^{oa}$, where '$a$' = '$a^0 a^1 a^2 \ldots a^m$' is the information contained in (b).

The general procedure is to minimise $I_\alpha$ with respect to $p_i^{oa}$ such that

$$ \sum_i p_i^{oa} a^k(x_i) = \langle a^k \rangle = 0 \,, \qquad (m+1 \text{ equations}) \,. \tag{23} $$

Form the Langrangian

$$ \mathcal{L} = I_\alpha + \sum_i \sum_k \lambda^k [p_i^{oa} a^k(x_i) - \langle a^k \rangle] \,, \tag{24} $$

which gives a minimum when

$$ \frac{\partial \mathcal{L}}{\partial p_i^{oa}} = \frac{\partial I_\alpha}{\partial p_i^{oa}} + \sum_k \lambda^k a^k(x_i) = 0 \,. \tag{25} $$

For the Taneja form we have

$$ \frac{\partial {}_T I_\alpha}{\partial p_i^{oa}} = \alpha^{-1}(2^{\alpha-1} - 1)(p_i^{oa})^{\alpha-1}(p_i^o)^{1-\alpha} \,, \tag{26} $$

so that

$$ p_i^{oa} = p_i^o [\alpha^{-1}(2^{\alpha-1} - 1) \sum_k \lambda^k a^k(x_i)]^\beta \,, \tag{27} $$

where $\beta = (\alpha - 1)^{-1}$ for all finite values of $\beta$ for which $\alpha > 0$.

For the Renyi proposal we have

$$\frac{\partial_R I_\alpha}{\partial p_i^{oa}} = U^{-1}\alpha^{-1}(\alpha-1)(p_i^{oa})^{\alpha-1}(p_i^o)^{1-\alpha} , \tag{28}$$

where $U = \sum_i (p_i^{oa})^{\alpha-1}(p_i^o)^{1-\alpha}$, so that

$$p_i^{oa} = p_i^o[U^{-1}\alpha^{-1}(\alpha-1)\sum_k \lambda^k a^k(x_i)]^\beta . \tag{29}$$

Both equations may be written in the form

$$p_i^{oa} = p_i^o\left[Q^{-1}\sum_k \lambda^k a^k(x_i)\right]^\beta , \tag{30}$$

where $Q^{-1}$ is some multiplier. From the constraint

$$\sum_i p_i^{oa} = 1 ,$$

we have

$$\sum_i p_i^{oa} = 1 = \sum_i p_i^o\left\{Q^{-1}\left[\lambda^o + \sum_{k>0}\lambda^k a^k(x_i)\right]\right\}^\beta , \tag{31}$$

$$= (\lambda^o)^\beta \sum_i p_i^o\left\{Q^{-1}\left[1 + \sum_{k>0}\nu^k a^k(x_i)\right]\right\}^\beta , \tag{32}$$

where $\lambda^o\nu^k = \lambda^k$. Hence

$$(\lambda^o)^\beta = \left[\left[\sum_i p_i^o\left\{Q^{-1}\left[1+\sum_{k>0}\nu^k a^k(x_i)\right]\right\}^\beta\right]\right]^{-1} . \tag{33}$$

Thus

$$p_i^{oa} = \frac{p_i^o\left\{Q^{-1}\left[1+\sum_{k>0}\nu^k a^k(x_i)\right]\right\}^\beta}{\sum_i p_i^o\left\{Q^{-1}\left[1+\sum_{k>0}\nu^k a^k(x_i)\right]\right\}^\beta} \tag{34}$$

$$= p_i^o Z^{-1}\left[1+\sum_{k>0}\nu^k a^k(x_i)\right]^\beta , \tag{35}$$

where $Z$ is the denominator in equation (34) multiplied by $Q$.

For Kullback/Hobson,

$$\frac{\partial_{KH} I}{\partial p_i^{oa}} = (\ln 2)^{-1}\left(\ln\frac{p_i^{oa}}{p_i^o} + 1\right) , \tag{36}$$

and

$$p_i^{oa} = p_i^o(\ln 2)\exp\left[\sum_k \lambda^k a^k(x_i)\right]$$

$$= p_i^o Z^{-1}\exp\left[\sum_{k>0}\lambda^k a^k(x_i)\right] , \tag{37}$$

where

$$Z = \sum_i p_i^o \exp\left[\sum_{k>0}\lambda^k a^k(x_i)\right] , \tag{38}$$

by applying the zeroth constraint and finding $\lambda^o$.

Equation (38) reduces to Jaynes' standard form when the $p_i^o = 1/n$, $\forall i$, that is, when the prior distribution is rectangular.

## Conclusion

Thus for Taneja and Renyi, $\alpha \neq 1$, we have the common form (35),

$$p_i^{oa} = p_i^o Z^{-1} \left[1 + \sum_{k>0} \nu^k a^k(x_i)\right]^{\beta} ,$$

for finite values of $\beta$; and for Kullback/Hobson the form (38),

$$p_i^{oa} = p_i^o Z^{-1} \exp\left[\sum_{k>0} \lambda^k a^k(x_i)\right] ,$$

as $\beta$ becomes infinitely large, which may be compared with Bayes' expression (1),

$$p_i^{oa} = p_i^o L^i .$$

It will be amply clear that the procedure of information minimizing estimates the likelihood ratio, $L^i$; it does not—as Jaynes suggests—estimate the prior probability. The confusion arises because Shannon's information implicitly assumes equal priors, $p_i^o = 1/n$, $\forall i$. Further there is an infinite number of such estimates depending on the choice of a value for $\alpha$, so that the uniqueness claimed for Jaynes' result is called into question too. If this infinity of estimates is accepted, then it would follow that the classic Gibbs/Boltzmann equation in thermodynamics for a perfect gas is but a particular case of an infinite number of equally acceptable models!

What then is special about the limit $\alpha \to 1$? This is a question which we address elsewhere (March and Batty, 1975b), but meanwhile the reader might be interested in papers by Aczél *et al.* (1974), and Tribus and Rossi (1973) where excellent arguments in favour of the uniqueness of Shannon's form are presented.

Although strong arguments can be brought forward to support the original Shannon-based formalism of Jaynes, we believe that our derivation from Bayes' theorem does present a new view on the subject, and that in design and planning, at least, there are sound pragmatic reasons for opting for other measures of information in model-building. In particular we have chosen the case $\alpha = 2$—corresponding directly to the ELR (3)—for the derivation of nonparametric forms for population and transportation models (March and Batty, 1975a), and in another paper we have discussed the rôle of prior probabilities in modelling spatial systems (Batty and March, 1975). We believe that such models may be justified.

This paper has demonstrated an information-theoretic approach to a central problem in the design activity: how to revise a proposal represented by $\{p_i^o\}$ given additional information and constraints '$a$'; that is, in the light of fresh hypotheses and new specifications. The revision represented by $\{p_i^{oa}\}$ is said to be least-prejudiced using $\alpha$-information measures. Of course, designers and planners may wish to assert their prejudices. Some may claim that design *is* the expression of personal taste. The approach taken here at least makes it possible—in theory—to measure how much *private* information a design reflects: a measure which may be of some use in assessing a designer's ability to meet a client's *publically* stated objectives, or *vice versa* to uncover hidden desires of the client as criticism of design proposals based on declared aims are made. In this way design becomes public.

### References

Aczél, J., Forte, B., Ng, C. T., 1974, "Why the Shannon and Hartley entropies are natural", *Advanced Applications in Probability*, **6**, 131-146.

Batty, M., 1974, "A theory of Markovian design machines", *Environment and Planning B*, **1**, 125-146.

Batty, M., March, L., 1975, "The method of residues in urban modelling", *The Transport Group Publication Series*, Department of Civil Engineering, University of Waterloo, Ontario, Canada.

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

Daróczy, A., 1970, "Generalized information functions", *Information and Control*, **16**, 36-51.
Evans, R. A., 1969, "The principle of minimum information", *IEEE Transactions on Reliability*, **18**, 87-90.
Feibleman, J. K., 1970, *An Introduction to the Philosophy of Charles S. Peirce* (MIT Press, Cambridge, Mass.).
Havdra, J., Charvát, F., 1967, "Quantification methods of classification processes: concepts of structural $\alpha$-entropy", *Kybernetika*, **3**, 30-34.
Hobson, A., Cheng, B. K., 1973, "A comparison of the Shannon and Kullback information measures", *Journal of Statistical Physics*, **7**, 302-310.
Jaynes, E. T., 1957, "Information theory and statistical mechanics" and "Information theory and statistical mechanics II", *Physical Review*, **106**, 620-630, **108**, 117-190.
Jaynes, E. T., 1968, "Prior probabilities", *IEEE Transactions on Systems, Man and Cybernetics*, **3**, 227-241.
Kullback, S., 1959, *Information Theory and Statistics* (John Wiley, New York).
March, L., 1975, "The logic of design and the question of value" in *The Architecture of Form* (Cambridge University Press, London). In the press.
March, L., Batty, M., 1975a, "Information-minimizing formalism and the derivation of non-parametric forms for population and transportation models", *Technical Report*, 6-S, Department of Systems Design, University of Waterloo, Ontario, Canada.
March, L., Batty, M., 1975b, "Measures of information and the derivation of statistical models", *Technical Report*, 7-S, Department of Systems Design, University of Waterloo, Ontario, Canada.
Renyi, A., 1961, "On measures of entropy and information", *Proceedings of the Fourth Symposium on Mathematical Statistics and Probability*, **1**, 547-561.
Shannon, C. F., 1948, "A mathematical theory of communication", *Bell System Technical Journal*, **27**, 379-423, 623-656.
Simon, H. A., 1969, *The Sciences of the Artificial* (MIT Press, Cambridge, Mass.).
Taneja, I. J., 1974, "A joint characterization of directed divergence, inaccuracy, and their generalizations", *Journal of Statistical Physics*, **11**, 169-176.
Tribus, M., 1969, *Rational Descriptions, Decisions and Designs* (Pergamon Press, New York).
Tribus, M., Rossi, R., 1973, "On the Kullback information measure as a basis for information theory: comments on a proposal by Hobson and Cheng", *Journal of Statistical Physics*, **9**, 331-338.