RAL-92-042

Science and Engineering Research Council
Rutherford Appleton Laboratory
Chilton DIDCOT Oxon OX11 0QX

RAL-92-042

The Information Content of Continuous Functions and the Geodesic Principle

M W Johnson

June 1992

The Information Content of Continuous Functions and the Geodesic Principle .

M.W.Johnson

Rutherford Appleton Laboratory
Chilton
Didcot
Oxon. OX11 0QX

June 1992

The Information Content of Continuous Functions and the Geodesic Principle

Abstract

The Maximum Entropy (ME) method is now widely used in the solution of inverse problems. This short paper re-iterates the fact that the classic ME method should only be applied to problems for which the solution is a probability distribution, and not to continuous functions of one or more variables. A method for quantifying the information content of continuous functions is given, and hence, through the Minimum Information principle, a technique for solving inverse problems where the solution is known to be continuous. The Geodesic Principle is introduced and *inter alia* we show why a straight line has the lowest information content of any continuous function.

Introduction

Inverse problems may be solved in many ways. Where the solution is known to be of a certain form that may be parameterised (e.g. atomic positions in crystal structure determination) the method of least squares may be used to establish the most likely model. Where a choice of model is difficult (in geological problems, for example) or uncertain, the required result is often a continuous function of one or more variables (the contours of geological strata, for example). In this case the problem is usually underdetermined and the inverse problem centres on finding the most acceptable model that is consistent with the available data. Since data are always noisy and incomplete the solution of inverse problems requires some method to distinguish between the (usually) infinite number of valid solutions that are consistent with the observed data.

A methodology, initiated by Jeffreys [1], has been developed over the last 30 years to deal with such problems. It is based on Bayes' theorem,

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} \qquad (1)$$

which relates the posterior probability distribution of hypotheses (H) given data (D), $P(H|D)$, to the prior probability function P(H) via the likelihood function $P(D|H)$. The calculation of the likelihood function is a direct problem, and is usually straightforward. The difficulty generally lies with the quantification of the prior knowledge, P(H), the probability distribution of the possible hypotheses.

The principle that is generally invoked is that of choosing the prior containing the least information - the non-informative prior. This is clearly related to Ockham's Razor [2]: given a series of possibilities that fit the facts one should chose the simplest.

In 1948 Shannon [3] showed how the "amount of uncertainty" in a probability distribution could be quantified, and derived a measure for the information content of a probability distribution. This measure is the expression;

$$I = -\sum_i p_i \log p_i \qquad (2)$$

and is the same as that for entropy in thermodynamics. Thus the technique of "Maximum Entropy" was developed, first by Jaynes[4], and subsequently by many others who have found it very effective in solving a variety of inverse problems [5].

Thus the ME technique consists of determining the probability distribution $\{p_i\}$ which maximises the entropy function:

$$\sum_i p_i \ln p_i \qquad (3)$$

while remaining consistent with the observed data.

However, as Jaynes points out [6] , Shannon's information measure (eq.2) is true only for discrete probability distributions, $\{p_i\}$. An example for which the method is entirely suitable would be the use of the ME method to decide between a number of alternative medical diagnoses.

Despite this apparent restriction, the ME method has been frequently applied to problems for which the final solution is a continuous function (f) or a discrete approximation to a continuous function $\{f_i\}$ referred to henceforth as a poly-line. (For examples see the papers listed under reference [5])

While apparently similar to a probability distribution (in that they are both n-tuples), the poly-line has the important distinction that it is an *ordered* n-tuple. Furthermore the polyline strictly requires n-(x,y) pairs for its description; the 'x' values for the function are often implied by the ordering parameter i in the y_i list. The concept of order does not enter into a probability distribution (p_i and p_j can be interchanged without affecting the solution, provided the items to which they refer are also interchanged) and hence the entropy function is invariant under a re-ordering of the set $\{p_i\}$.

Because of the insensitivity of the entropy expression to re-ordering, the entropy for the curve shown in Fig.(1a) is identical to that for the curve shown in Fig.(1b), which has been obtained by exchanging two pairs of y_i values. Intuitively it seems unreasonable that these two curves contain the same amount of information. Similarly, if the ME method is used to derive a smoothed fit to some experimental data, the tendency of the ME method to produce the 'most uniform' solution (i.e. all p_is equal) means that a 'horizontal' solution is preferred even when the data are consistent with an inclined straight line.

Recently Soper [7] pointed out the re-ordering paradox and examined the problem of inverting the neutron liquid structure factor, S(Q), to determine the atom pair correlation function, g(r). Here the problem is clearly one in which the desired solution is a continuous function, g(r), and he came to the conclusion that a more satisfactory result was obtained if *the least noisy* function, consistent with the data, was chosen. To do this the functional,

$$\sum_i w_i f_i''^2 \qquad (4)$$

was minimised, where f'' is the local second derivative and w a weight which de-emphasised the regions where the data was changing rapidly.

This is one of a number of 'minimum structure solutions' that have been used for solving inverse problems (see for example Aldridge et al [8]). Minimising the l_2 norm of f, f' or f'' (eq.4) produces the 'smallest', 'flattest' and 'smoothest' functions respectively. As will be seen, the choice of the 'smoothest' model is closely related to the results obtained below.

The proposition put forward in this paper is that a new measure is required for quantifying the information content of a continuous, or poly-line distribution, and this new measure should lead to improvements in the solution of inverse problems.

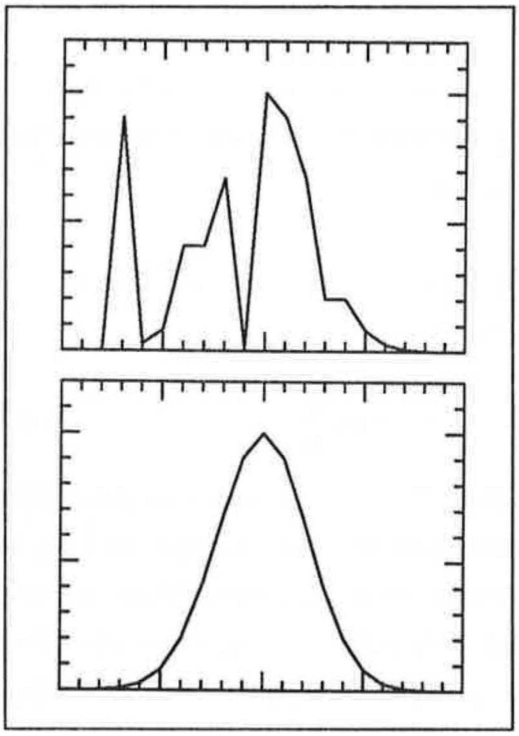

Fig.1a(upper), b(lower)

Continuous Functions

To understand how inverse problems should be solved when the solution is known to be a continuous function we must return to first principles. The principle behind the ME method is not entropy per se but *information*, and the ME formalism is one which selects the probability distribution which is both consistent with the data, and which contains the least information.

The first problem is therefore to establish the information content of a continuous function.

Shannon [3] defines the information in a message as:

$$I = K \ln \frac{N_0}{N_1} \qquad (5)$$

where N_1 are the number of possibilities after receiving the message and N_0 the number before. K simply defines the units, and if chosen to be $\log_2 e$ the information is in 'bits'. Consider first two simple messages;

$21^{o}00'$ N $40^{o}00'$ W. (m1)
21^{o} N 40^{o} W (m2)

The information in the two messages is the same but the information *content* is different. The information content is measured not from its actual content, but what the contents might have been. Thus the information content of (m1) is higher than that of (m2) since in the first case the number of possibilities is 2.3 10^8 (±90o in 1' steps and 0 - 360o in 1' steps) whereas in the second the number of possible messages is only 6.5 10^4 (1^{o} steps rather than 1'). The information contents of the two messages shown above are therefore

$$I1 = 1.443 \ln(2.3\ 10^8)$$
$$= 27.8 \text{ bits}$$

and

$$I2 = 1.442 \ln(6.5\ 10^4)$$
$$= 16.0 \text{ bits}$$

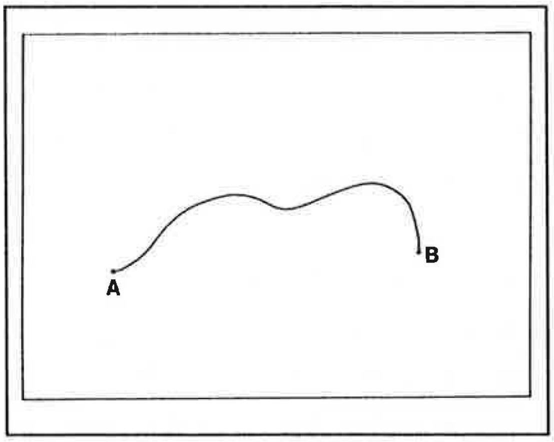

Fig.2

So what is the information content of the continuous curve shown in Fig.2 ? Clearly, this line is a message, in the general sense of the word, and in order to quantify its information content we must count the number of ways in which this line *could* have been drawn. This in turn obviously depends upon the 'resolution' with which the line is defined. If we define a distance r as the resolution length of the curved line which has a length l ($r << l$), we may then calculate the number of ways from simple random walk arguments. The number of ways N, that a line of length l can be drawn between two points a distance L apart is found to be (see Appendix) :

$$N = \exp\left\{ \frac{l}{r}\left(1 - L^2/l^2\right) \right\} \qquad (6)$$

This is the value of N_0 in eq.(5) and since N_1=1 the information content of such a line is:

$$I = K \frac{l}{r}\left(1 - L^2/l^2\right) \qquad (7)$$

It is therefore suggested that where the solution of an inverse problem is in the form of a continuous function it expression (7) which must be minimised subject to the constraints imposed by the data. The form of equation (7) is shown in Fig.3 (for L=1,r=0.01), illustrating the rapid asymptotic approach to the linear dependence of I on *l*.

Hence the curved line which contains the minimum information is the *shortest* line, and hence the solution of inverse problems with continuous solutions requires us to select the *shortest* curve consistent with the data. This may be succinctly referred to as the Geodesic Principle. The functional which must be minimised is therefore:

$$l = \int_A^B \left(1 + f'^2\right)^{1/2} dx \qquad (8)$$

Discussion

The above result explains why the use of the 'smoothest' solution (i.e. minimising l_2 norm of f'') achieves satisfactory results, since this also has the effect of minimising the length of the function. It also resolves several other points which were unsatisfactory when approached using the ME method.

If the ME method is used in a smoothing algorithm the result tends towards a horizontal straight line, since the entropy is maximised when all p_is are equal. This

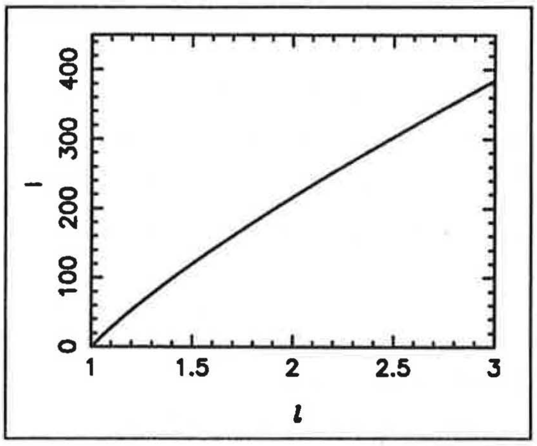

Fig.3

will be true even when the data are consistent with an inclined straight line. Using the Geodesic Principle the data will be correctly fitted by an inclined straight line.

It explains why the data in Fig.1a contains more information than that in Fig.1b, even though they have the same entropy. In addition, the information content of a continuous function, when measured using equation (7), is invariant under inversions, translations and rotations of the co-ordinate axes.

Most importantly the Geodesic Principle should introduce fewer problems when units are changed. Discrete probability distributions are unit-less (for example the probabilities of different weather patterns tomorrow) and the application of the ME method to such problems poses no difficulties. When the ME method is applied to spectra, for example, where the result is a (continuous) function, altering the units used for the x-axis produces a different result. This is again because of the tendency of the ME method to produce the most *'horizontal'* result, which clearly results in different solutions if the units of the abscissa are changed from u to $1/u$. The Geodesic Principle will tend to produce the *'straightest'* solution which will obviously introduce errors where the solution is curved, and this ,in turn, will depend on the units used. However, it is possible that this error might be allowed for in inversion algorithms.

It is also clear that the method will extend to two and more dimensions. While not a quantitative argument, it is straightforward to show that a larger surface may be arranged in more ways than a smaller one. (This can be seen by putting a 'loop' in the larger surface to reduce it to the same area as the smaller area.) Hence it is suggested that where the solution to an inverse problem is a surface or volume element, these also should be minimised in a manner similar to the Geodesic Principle.

Acknowledgements

I would like to thank Bill David, Alan Soper and particularly Devinder Sivia for many illuminating discussions on inverse problems and 'logical inference'.

Appendix

We consider here the 2-d case , although the generalisation to 3 or more dimensions would appear possible. The problem is to calculate the number of possible continuous functions that join the points A,B in Fig.2. We represent the line by a poly-line containing n segments each of a length r, the 'resolution length' used in the description of the curve. Hence

$$l = n.r \qquad \text{(A1)}$$

and $$r << l \; .$$

We consider how many ways a n-polyline of length l might be arranged with its starting point at the origin, A, and no restriction placed on its other end B. The problem is simply that of a fixed-step random walk. The probability distribution for the x projection of each step is,

$$\begin{aligned} P(x) &= 1/\pi\,(r^2 - x^2)^{1/2} \quad & |x|\leq r \qquad \text{(A2)} \\ &= 0 & |x|>r \end{aligned}$$

which has a mean of 0, and a variance $\sigma^2 = r^2/2$. The addition of each of the random steps results in a convolution of their probability distributions which, from the central limit theorem, is a Gaussian distribution. Thus the probability distributions of the x and y co-ordinates of the end B, are Gaussian distributions centred at 0, with variances $\sigma^2 = nr^2/2$ (second moments added in quadrature, a result which holds even for non-Gaussian distribution functions). Thus the probability of the n-polyline ending at a point (x,y) per unit area is;

$$P(x,y) = \frac{1}{\pi n r^2} e^{-(x^2+y^2)/nr^2} \qquad \text{(A3)}$$

Denoting the total distance A-B as L and noting $nr = l$ we may rewrite (A3) as

$$P(L) = \frac{1}{\pi r l} e^{-L^2/rl} \qquad \text{(A4)}$$

Let us further assume that the number of possible angular positions of each step is Θ and the definition of the point B is an area of size r^2 (i.e. dimensions similar to that of the resolution of the polyline).
The *number* of ways in which the polyline can finish at B is therefore,

$$N(L) = \frac{\Theta^n e^{-L^2/rl}}{\pi n} \qquad \text{(A5)}$$

This equation is clearly an approximation since it overstates the number when L > l , which is ,of course, zero. However we may use the fact that N(L) = 1 when L = l to derive a 'normalising' value for Θ, which is

$$\Theta = e(\pi n)^{1/n} \qquad \text{(A6)}$$

Substituting (A6) in (A5) gives

$$N = e^{n(1-L^2/l^2)} \qquad \text{(A7)}$$

or

$$N = e^{l(1-L^2/l^2)/r} \qquad \text{(A8)}$$

References

1. 'Theory of probability' H. Jeffreys O.U.P. (1939)
2. see for example ref.1, p.342
3. 'The mathematical Theory of Communication' C.E.Shannon & W.Weaver University of Illinois Press. (1963)
4. 'E.T.Jaynes:Papers on Probability, Statistics and Statistical Physics' Ed. R.D.Rosenkrantz D.Reidel (1983)
5. See the conference reports 'Maximum Entropy and Bayesian Methods' in the series Fundamental Theories of Physics, Kluwer (1989)
6. p.59 of reference 4
7. 'Neutron Scattering Data Analysis' Ed. M.W.Johnson, Institute of Physics Conference Series 107 IOP (1990)
8. 'New Methods for Constructing Flattest and Smoothest models' D.F.Aldridge et al Inverse Problems **7** 499-513 (1991)